Physics

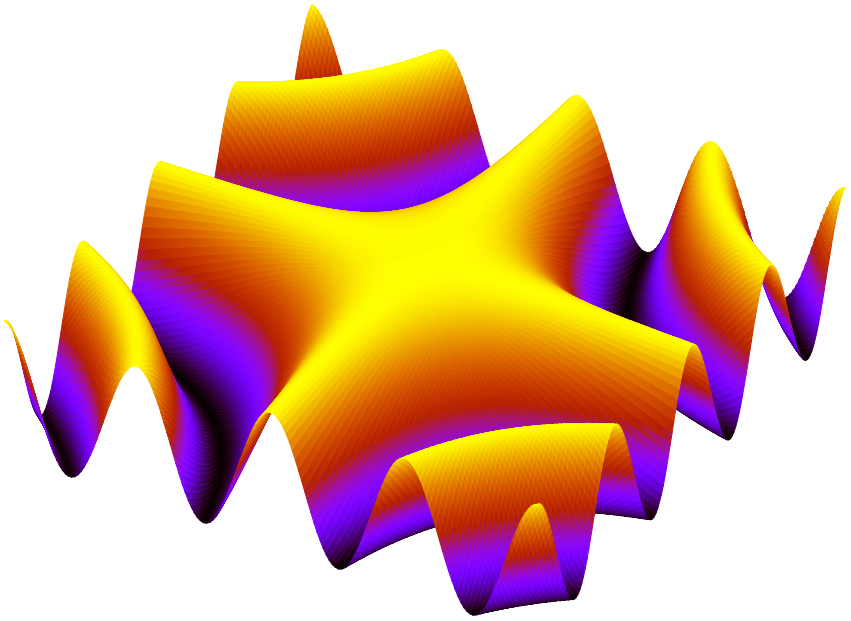
with a
Strange Sense of Humor,
Some Moralizing,
and a
Few Duly Cynical Observations
About Humanity

John Gaffney

Draft Edition of 7 June 2019

Copyright © 2003-2019 John Gaffney

This book is a work of fiction.
Any truths or resemblance to reality are entirely coincidental.



This completely gratuitous plot of $z = \cos(xy)$
has absolutely nothing to do with anything.

$$\int_1^{\sqrt[3]{3}} z^2\, dz \cdot \cos\frac{3\pi}{9} = \ln\sqrt[3]{e}$$

Neither does the above limerick (author unknown).

This work is licensed under the Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 United States License. To view a copy of this license, visit

http://creativecommons.org/licenses/by-nc-nd/3.0/us/legalcode

License

THE WORK (AS DEFINED BELOW) IS PROVIDED UNDER THE TERMS OF THIS CREATIVE COMMONS PUBLIC LICENSE ("CCPL" OR "LICENSE"). THE WORK IS PROTECTED BY COPYRIGHT AND/OR OTHER APPLICABLE LAW. ANY USE OF THE WORK OTHER THAN AS AUTHORIZED UNDER THIS LICENSE OR COPYRIGHT LAW IS PROHIBITED.

BY EXERCISING ANY RIGHTS TO THE WORK PROVIDED HERE, YOU ACCEPT AND AGREE TO BE BOUND BY THE TERMS OF THIS LICENSE. TO THE EXTENT THIS LICENSE MAY BE CONSIDERED TO BE A CONTRACT, THE LICENSOR GRANTS YOU THE RIGHTS CONTAINED HERE IN CONSIDERATION OF YOUR ACCEPTANCE OF SUCH TERMS AND CONDITIONS.

1. Definitions

    a. "Collective Work" means a work, such as a periodical issue, anthology or encyclopedia, in which the Work in its entirety in unmodified form, along with one or more other contributions, constituting separate and independent works in themselves, are assembled into a collective whole. A work that constitutes a Collective Work will not be considered a Derivative Work (as defined below) for the purposes of this License.

    b. "Derivative Work" means a work based upon the Work or upon the Work and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Work may be recast, transformed, or adapted, except that a work that constitutes a Collective Work will not be considered a Derivative Work for the purpose of this License. For the avoidance of doubt, where the Work is a musical composition or sound recording, the synchronization of the Work in timed-relation with a moving image ("synching") will be considered a Derivative Work for the purpose of this License.

    c. "Licensor" means the individual, individuals, entity or entities that offers the Work under the terms of this License.

    d. "Original Author" means the individual, individuals, entity or entities who created the Work.

    e. "Work" means the copyrightable work of authorship offered under the terms of this License.

    f. "You" means an individual or entity exercising rights under this License who has not previously violated the terms of this License with respect to the Work, or who has received express permission from the Licensor to exercise rights under this License despite a previous violation.

2. Fair Use Rights. Nothing in this license is intended to reduce, limit, or restrict any rights arising from fair use, first sale or other limitations on the exclusive rights of the copyright owner under copyright law or other applicable laws.

3. License Grant. Subject to the terms and conditions of this License, Licensor hereby grants You a worldwide, royalty-free, non-exclusive, perpetual (for the duration of the applicable copyright) license to exercise the rights in the Work as stated below:

    a. to reproduce the Work, to incorporate the Work into one or more Collective Works, and to reproduce the Work as incorporated in the Collective Works; and,

    b. to distribute copies or phonorecords of, display publicly, perform publicly, and perform publicly by means of a digital audio transmission the Work including as incorporated in Collective Works.

    The above rights may be exercised in all media and formats whether now known or hereafter devised. The above rights include the right to make such modifications as are technically necessary to exercise the rights in other media and formats, but otherwise you have no rights to make Derivative Works. All rights not expressly granted by Licensor are hereby reserved, including but not limited to the rights set forth in Sections 4(d) and 4(e).

4. Restrictions.The license granted in Section 3 above is expressly made subject to and limited by the following restrictions:

    a. You may distribute, publicly display, publicly perform, or publicly digitally perform the Work only under the terms of this License, and You must include a copy of, or the Uniform Resource Identifier for, this License with every copy or phonorecord of the Work You distribute, publicly display, publicly perform, or publicly digitally perform. You may not offer or impose any terms on the Work that restrict the terms of this License or the ability of a recipient of the Work to exercise the rights granted to that recipient under the terms of the License. You may not sublicense the Work. You must keep intact all notices that refer to this License and to the disclaimer of warranties. When You distribute, publicly display, publicly perform, or publicly digitally perform the Work, You may not impose any technological measures on the Work that restrict the ability of a recipient of the Work from You to exercise the rights granted to that recipient under the terms of the License. This Section 4(a) applies to the Work as incorporated in a Collective Work, but this does not require the Collective Work apart from the Work itself to be made subject to the terms of this License. If You create a Collective Work, upon notice from any Licensor You must, to the extent practicable, remove from the Collective Work any credit as required by Section 4(c), as requested.

    b. You may not exercise any of the rights granted to You in Section 3 above in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation. The exchange of the Work for other copyrighted works by means of digital file-sharing or otherwise shall not be considered to be intended for or directed toward commercial advantage or private monetary compensation, provided there is no payment of any monetary compensation in connection with the exchange of copyrighted works.

    c. If You distribute, publicly display, publicly perform, or publicly digitally perform the Work (as defined in Section 1 above) or Collective Works (as defined in Section 1 above), You must, unless a request has been made pursuant to Section 4(a), keep intact all copyright notices for the Work and provide, reasonable to the medium or means You are utilizing: (i) the name of the Original Author (or pseudonym, if applicable) if supplied, and/or (ii) if the Original

Author and/or Licensor designate another party or parties (e.g. a sponsor institute, publishing entity, journal) for attribution ("Attribution Parties") in Licensor's copyright notice, terms of service or by other reasonable means, the name of such party or parties; the title of the Work if supplied; to the extent reasonably practicable, the Uniform Resource Identifier, if any, that Licensor specifies to be associated with the Work, unless such URI does not refer to the copyright notice or licensing information for the Work. The credit required by this Section 4(c) may be implemented in any reasonable manner; provided, however, that in the case of a Collective Work, at a minimum such credit will appear, if a credit for all contributing authors of the Collective Work appears, then as part of these credits and in a manner at least as prominent as the credits for the other contributing authors. For the avoidance of doubt, You may only use the credit required by this clause for the purpose of attribution in the manner set out above and, by exercising Your rights under this License, You may not implicitly or explicitly assert or imply any connection with, sponsorship or endorsement by the Original Author, Licensor and/or Attribution Parties, as appropriate, of You or Your use of the Work, without the separate, express prior written permission of the Original Author, Licensor and/or Attribution Parties.

d. For the avoidance of doubt, where the Work is a musical composition:

   i. Performance Royalties Under Blanket Licenses. Licensor reserves the exclusive right to collect whether individually or, in the event that Licensor is a member of a performance rights society (e.g. ASCAP, BMI, SESAC), via that society, royalties for the public performance or public digital performance (e.g. webcast) of the Work if that performance is primarily intended for or directed toward commercial advantage or private monetary compensation.

   ii. Mechanical Rights and Statutory Royalties. Licensor reserves the exclusive right to collect, whether individually or via a music rights agency or designated agent (e.g. Harry Fox Agency), royalties for any phonorecord You create from the Work ("cover version") and distribute, subject to the compulsory license created by 17 USC Section 115 of the US Copyright Act (or the equivalent in other jurisdictions), if Your distribution of such cover version is primarily intended for or directed toward commercial advantage or private monetary compensation.

e. Webcasting Rights and Statutory Royalties. For the avoidance of doubt, where the Work is a sound recording, Licensor reserves the exclusive right to collect, whether individually or via a performance-rights society (e.g. SoundExchange), royalties for the public digital performance (e.g. webcast) of the Work, subject to the compulsory license created by 17 USC Section 114 of the US Copyright Act (or the equivalent in other jurisdictions), if Your public digital performance is primarily intended for or directed toward commercial advantage or private monetary compensation.

5. Representations, Warranties and Disclaimer

   UNLESS OTHERWISE MUTUALLY AGREED TO BY THE PARTIES IN WRITING, LICENSOR OFFERS THE WORK AS-IS AND ONLY TO THE EXTENT OF ANY RIGHTS HELD IN THE LICENSED WORK BY THE LICENSOR. THE LICENSOR MAKES NO REPRESENTATIONS OR WARRANTIES OF ANY KIND CONCERNING THE WORK, EXPRESS, IMPLIED, STATUTORY

OR OTHERWISE, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF TITLE, MARKETABILITY, MERCHANTIBILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR THE ABSENCE OF LATENT OR OTHER DEFECTS, ACCURACY, OR THE PRESENCE OF ABSENCE OF ERRORS, WHETHER OR NOT DISCOVERABLE. SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OF IMPLIED WARRANTIES, SO SUCH EXCLUSION MAY NOT APPLY TO YOU.

6. Limitation on Liability. EXCEPT TO THE EXTENT REQUIRED BY APPLICABLE LAW, IN NO EVENT WILL LICENSOR BE LIABLE TO YOU ON ANY LEGAL THEORY FOR ANY SPECIAL, INCIDENTAL, CONSEQUENTIAL, PUNITIVE OR EXEMPLARY DAMAGES ARISING OUT OF THIS LICENSE OR THE USE OF THE WORK, EVEN IF LICENSOR HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

7. Termination

    a. This License and the rights granted hereunder will terminate automatically upon any breach by You of the terms of this License. Individuals or entities who have received Collective Works (as defined in Section 1 above) from You under this License, however, will not have their licenses terminated provided such individuals or entities remain in full compliance with those licenses. Sections 1, 2, 5, 6, 7, and 8 will survive any termination of this License.

    b. Subject to the above terms and conditions, the license granted here is perpetual (for the duration of the applicable copyright in the Work). Notwithstanding the above, Licensor reserves the right to release the Work under different license terms or to stop distributing the Work at any time; provided, however that any such election will not serve to withdraw this License (or any other license that has been, or is required to be, granted under the terms of this License), and this License will continue in full force and effect unless terminated as stated above.

8. Miscellaneous

    a. Each time You distribute or publicly digitally perform the Work (as defined in Section 1 above) or a Collective Work (as defined in Section 1 above), the Licensor offers to the recipient a license to the Work on the same terms and conditions as the license granted to You under this License.

    b. If any provision of this License is invalid or unenforceable under applicable law, it shall not affect the validity or enforceability of the remainder of the terms of this License, and without further action by the parties to this agreement, such provision shall be reformed to the minimum extent necessary to make such provision valid and enforceable.

    c. No term or provision of this License shall be deemed waived and no breach consented to unless such waiver or consent shall be in writing and signed by the party to be charged with such waiver or consent.

    d. This License constitutes the entire agreement between the parties with respect to the Work licensed here. There are no understandings, agreements or representations with respect to the Work not specified here. Licensor shall not be bound by any additional provisions that may appear in any communication from You. This License may not be modified without the mutual written agreement of the Licensor and You.

Contents

License **iii**

Other Stuff **xvii**

I Preliminaries 1

–1 Preliminaries **3**
- –1.1 About the Course . . . 3
- –1.2 About This Book . . . 3
- –1.3 Good Karma . . . 5
- –1.4 The Zen of Problem Solving . . . 7
  - –1.4.1 An Unfortunate Example . . . 11
- –1.5 Significant Figures . . . 16
- –1.6 Units & Conversions . . . 19
- –1.7 Conversion Factors & Constants . . . 19
- –1.8 Order-of-Magnitude Estimates . . . 21
  - –1.8.1 An Example . . . 21
  - –1.8.2 General Points . . . 23
  - –1.8.3 A Brief Discourse on Malarkey . . . 24
- –1.9 Problems . . . 25
- –1.10 Sketchy Answers . . . 27

0 Optics **29**
- 0.1 Light Waves . . . 29
- 0.2 Geometrical Optics . . . 32
  - 0.2.1 Reflection & Refraction . . . 32
  - 0.2.2 Ray Diagrams & Images . . . 35
  - 0.2.3 Thin Lenses . . . 37
  - 0.2.4 Optical Instruments . . . 42
  - 0.2.5 The Eye & Corrective Lenses . . . 43

0.3 Wave Optics . . . 45
0.3.1 Interference, Diffraction, Dispersion, & Polarization . . 45
0.3.2 Why the Sky is Blue . . . 49
0.4 Parabolic Mirrors . . . 52
0.5 Problems . . . 59
0.6 Sketchy Answers . . . 64

1 Vectors **65**
1.1 Unit Vectors . . . 69
1.2 Dot & Cross Products . . . 71
1.2.1 The Dot Product . . . 71
1.2.2 The Cross Product . . . 73
1.2.3 Some Special Cases . . . 78
1.3 Problems . . . 79
1.4 Sketchy Answers . . . 83

2 Vector Calculus **85**
2.1 Line Elements & Integrations . . . 85
2.2 Surface Elements & Integrations . . . 90
2.3 Volume Elements & Integrations . . . 92
2.4 The Gradient . . . 94
2.5 Divergence & Gauss's Theorem . . . 96
2.6 Curl & Stokes's Theorem . . . 100
2.6.1 An Important Result . . . 106
2.7 A Few More Important Results . . . 107
2.7.1 $\nabla \cdot \mathbf{B} = 0 \Rightarrow \mathbf{B} = \nabla \times \mathbf{A}$ . . . 107
2.7.2 The Behavior of $\nabla^2 \frac{1}{r}$ . . . 109
2.7.3 Helmholtz's Theorem . . . 112
2.8 Problems . . . 115
2.9 Sketchy Answers . . . 119

II Basic Mechanics **121**

3 Kinematics **123**
3.1 Location, Velocity, & Acceleration . . . 123
3.2 One-Dimensional Motion . . . 129
3.2.1 Constant Acceleration . . . 132
3.2.2 Vertical Free-Fall . . . 135
3.3 Two-Dimensional Motion . . . 136
3.3.1 Projectile Motion . . . 136
3.3.2 Uniform Circular Motion . . . 139
3.3.3 Centripetal Acceleration, Newton-Style . . . 144

3.3.4 Nonuniform Circular Motion . . . 146
3.3.5 General Motion in Polar Coordinates . . . 149
3.3.6 Two-Dimensional Relative Velocities . . . 153
3.4 Problems . . . 154
3.5 Sketchy Answers . . . 185

4 Dynamics 191
4.1 Newton's Laws . . . 191
4.2 Special Forces . . . 194
4.3 Force Diagrams . . . 197
4.4 Circular Motion . . . 204
4.4.1 Road Banking . . . 207
4.5 Newton's Law of Gravity & Orbits . . . 208
4.6 Perceived Weight . . . 212
4.7 Semi- & Almost Nonbogus Friction . . . 214
4.8 The Catenary . . . 216
4.9 Problems . . . 221
4.10 Sketchy Answers . . . 249

5 Work & Energy 253
5.1 The Bogonics of Work & Power . . . 254
5.2 Potential Energy & Energy Conservation . . . 260
5.3 A Practical Example of Energy Conservation . . . 264
5.4 Results for Potential Energy . . . 266
5.5 How to Beat a Dead Horse . . . 269
5.6 Problems . . . 272
5.7 Sketchy Answers . . . 289

6 Center of Mass & Momentum 291
6.1 Center of Mass . . . 291
6.2 The Dynamics of the Center of Mass . . . 298
6.3 Momentum & Momentum Conservation . . . 299
6.4 Collisions . . . 302
6.5 Center-of-Mass & Relative Coordinates . . . 303
6.6 Rockets . . . 306
6.7 Two-Body Collisions . . . 307
6.7.1 The One-Dimensional Two-Body Elastic Collision . . . 311
6.8 Summary of Important Points . . . 314
6.9 Some Gravitational Yawing . . . 317
6.10 Problems . . . 324
6.11 Sketchy Answers . . . 344

7 Rotational Dynamics **347**
7.1 Two-Dimensional Rotations . . . 347
Table of Translational-Rotational Analogies . . . 356
7.2 Three-Dimensional Rotations . . . 357
Weird Properties of Three-Dimensional Rotations . . . 364
7.3 Coriolis Effects . . . 366
7.4 Constant Angular Acceleration . . . 373
7.5 Moments of Inertia . . . 374
7.6 The Parallel-Axis Theorem . . . 379
7.7 Conservation of Angular Momentum . . . 381
7.8 Kinetic Energy . . . 390
7.9 Torque Due to Gravity . . . 392
7.10 Rolling . . . 393
7.10.1 All Good Things Must Come to an End . . . 399
7.11 Massive Pulleys . . . 400
7.12 Gyroscopes & Tops . . . 401
7.13 Summary of Important Points . . . 403
7.14 Problems . . . 406
7.15 Sketchy Answers . . . 436

8 Static Equilibria **441**
8.1 The Conditions of Equilibrium . . . 441
8.2 Stable & Unstable Equilibria . . . 445
8.3 Problems . . . 448
8.4 Sketchy Answers . . . 457

9 Harmonic Motion **459**
9.1 The First Refrain . . . 459
9.2 Once Again, with Feeling . . . 465
9.3 Some Practical Considerations . . . 468
9.4 Pendula . . . 470
9.5 Damped Harmonic Oscillations . . . 477
9.6 Driven Damped Harmonic Oscillations . . . 479
9.7 Small Oscillations . . . 484
9.8 Wave Effects . . . 486
9.8.1 Traveling Waves . . . 486
9.8.2 Standing Waves . . . 487
9.8.3 Sound Intensity & Decibels . . . 489
9.8.4 Doppler Shift . . . 490
9.9 Problems . . . 493
9.10 Sketchy Answers . . . 508

III Beyond Basic Mechanics 509

10 Relativity 511
10.1 Reference Frames . . . 511
10.2 Einstein Disses Newton . . . 513
10.3 The Lorentz Transform . . . 517
10.3.1 Special Case: A Mirror & Light Pulse . . . 518
10.3.2 Derivation of the Lorentz Transform . . . 521
10.3.3 A Nicer Derivation of the Lorentz Transform . . . 525
10.3.4 A More Modern Derivation of the Lorentz Transform 528
10.3.5 Some Observations & Notation . . . 532
10.3.6 The Inverse Lorentz Transform . . . 536
10.3.7 The Lorentz Transform from Symmetry . . . 538
10.3.8 Lorentz Transforms as Rotations . . . 549
10.4 Time Dilation & Length Contraction . . . 555
10.4.1 Time Dilation . . . 555
10.4.2 Length Contraction . . . 558
10.4.3 When to Use What . . . 561
10.4.4 One Final Note . . . 562
10.5 The Invariant Interval & Proper Time . . . 563
10.6 Addition of Velocities . . . 567
10.7 Momentum, Energy, & Stuff . . . 569
10.7.1 A Nicer Derivation of Momentum & Energy . . . 575
10.8 The Doppler Shift . . . 577
10.9 General Relativity . . . 579
10.9.1 The Field Equations . . . 583
10.9.2 Gravitational Time Dilation . . . 587
10.10 Constant Acceleration . . . 588
10.11 Problems . . . 599
10.12 Sketchy Answers . . . 618

11 Fluid Dynamics 621
11.1 The Bernoulli Equation . . . 621
11.2 Archimedes's "Principle" . . . 624
11.3 Frisbees & Airplanes . . . 625
11.4 Brazilian Soccer . . . 626
11.5 Why Golf Balls Have Dimples . . . 627
11.6 Problems . . . 629
11.7 Sketchy Answers . . . 635

12 Things Nuclear 637
12.1 The Composition of Nuclei . . . 637
12.2 Types of Nuclear Decay . . . 639
12.3 Decay Rates & Constants . . . 641
12.4 Fission & Fusion . . . 644

12.5 Dosimetry & Biological Effects 648
12.6 One More Reason New Jersey Is Disgusting 656
12.7 Problems 658
12.8 Sketchy Answers 662
Periodic Table 663

13 Thermal Physics **665**
13.1 Statistical Mechanics Versus Thermodynamics 665
13.2 Some (*ugh!*) Chemistry 669
13.3 Temperature Scales 670
13.4 Heat Energy & Changes of Temperature & Phase 671
13.5 Ideal Gases 676
13.6 Processes, Cycles, & the First Law 685
13.6.1 A Painfully Long Example 688
13.6.2 A Mercifully Short Example 694
13.6.3 Adiabatic Processes 697
13.6.4 Some General Observations 702
13.7 Heat Engines, & Refrigerators 704
13.7.1 The Carnot Cycle 704
13.7.2 Air Conditioning & Refrigeration 714
13.8 Reversibility, Entropy, & the Second Law 717
13.8.1 Entropy in Statistical Mechanics 725
13.8.2 Some Examples & Observations 729
13.9 Problems 734
13.10 Sketchy Answers 744

IV Electromagnetism for Big People **747**

14 The Maxwell Equations: An Overview **749**
14.1 The Maxwell Equations 749
14.2 Charge & Current 750
14.3 Gauss's Law & Electric Fields & Forces 753
14.4 Magnetic Gauss's Law & Magnetic Fields 755
14.5 Faraday's Law 756
14.6 Ampère's Law & Magnetic Forces 758
14.7 Superposition 761
14.8 The Potential Functions 762
14.8.1 Gauge Transforms & Gauge Symmetry 764
14.9 Light, Locality, & Relativity 765

15 Electrostatics **773**
15.1 Applications of Gauss's Law . . . 773
15.1.1 Spherical Charge Distributions . . . 774
15.1.2 Cylindrical Charge Distributions . . . 778
15.1.3 Planar Charge Distributions . . . 780
15.1.4 Superposition . . . 783
15.2 Coulomb's Semibogus Law . . . 784
15.3 Electric Fields by Direct Integration . . . 786
15.3.1 Rings of Charge . . . 786
15.3.2 Disks of Charge . . . 788
15.3.3 Finite Line Segments of Charge . . . 791
15.4 Electric Field Lines . . . 794
15.5 Electric Dipoles . . . 797
15.6 Electrostatic Potential & Voltage . . . 799
15.7 Equipotential Lines & Surfaces . . . 802
15.8 Electrostatic Potential Energy . . . 807
15.9 Conductors . . . 809
15.10 The Method of Images . . . 812
15.11 Problems . . . 815
15.12 Sketchy Answers . . . 835

16 DC Circuits **839**
16.1 Resistance & Power . . . 839
16.2 Series & Parallel Connections . . . 842
16.3 Loop & Junction Rules . . . 849
16.4 Capacitance . . . 852
16.4.1 Parallel-Plate Capacitors . . . 852
16.4.2 Cylindrical Capacitors . . . 853
16.4.3 Spherical Capacitors . . . 854
16.4.4 A Few Observations . . . 856
16.4.5 Dielectrics . . . 857
16.5 Capacitors in Circuits . . . 858
16.5.1 Energy Stored in a Capacitor . . . 861
16.5.2 Games People Play . . . 862
16.6 *RC* Circuits . . . 865
16.7 Treating AC As DC . . . 868
16.8 Problems . . . 871
16.9 Sketchy Answers . . . 886

17 Magnetostatics **889**
17.1 Magnetic Forces . . . 891
17.2 Ampère's Law . . . 893
17.2.1 Field of an Infinite Wire . . . 893
17.2.2 Field of a Solenoid . . . 897

17.3 The Biot-Savart Law . . . 901
17.3.1 Derivation of the Biot-Savart Law . . . 901
17.3.2 Field of an Infinite Wire . . . 906
17.3.3 Field of a Circular Arc . . . 907
17.3.4 Field of a Ring of Current . . . 908
17.3.5 Field of a Solenoid . . . 909
17.4 Magnetic Dipoles . . . 910
17.5 Problems . . . 915
17.6 Sketchy Answers . . . 925

18 Electrodynamics 927
18.1 Faraday's Law . . . 927
18.2 Ye Olde Sliding Bar . . . 935
18.3 Generators & Motors . . . 937
18.4 On the Issue of Time Derivatives . . . 939
18.5 Problems . . . 944
18.6 Sketchy Answers . . . 952

19 More DC Circuits 953
19.1 Inductance . . . 953
19.2 *LR* Circuits . . . 956
19.3 Energy Density of the Electromagnetic Field . . . 958
19.4 Problems . . . 963
19.5 Sketchy Answers . . . 967

20 AC Circuits 969
20.1 Fourier Transforms . . . 969
20.2 Impedance . . . 978
20.2.1 The *RC* Circuit . . . 981
20.2.2 The *LR* Circuit . . . 985
20.2.3 The *RLC* Circuit . . . 985
20.3 Delta Functions for Dummies . . . 988

V And Now for Something Completely Different . . . 991

21 Lagrangian Dynamics 993
21.1 The Calculus of Variations . . . 993
21.2 The Brachistochrone . . . 997
21.2.1 A Brief Digression, For Those Inclined to Digress . . . 1002
21.3 Lagrangian Dynamics . . . 1005
21.4 Lagrange Multipliers & Constraints . . . 1011
21.5 Forces of Constraint . . . 1015
21.6 Problems . . . 1020
21.7 Sketchy Answers . . . 1033

22 Real Physics **1035**
22.1 The Early Days . . . 1038
22.2 Newton . . . 1039
22.3 Maxwell & Others . . . 1039
22.4 Relativity . . . 1041
22.5 Quantum Mechanics . . . 1042
22.5.1 Wave Functions & Operators . . . 1047
22.5.2 The Schrödinger Equation & Discrete States . . . 1052
22.5.3 Quantum Tunneling . . . 1058
22.5.4 The Quantum Harmonic Oscillator . . . 1059
22.5.5 Path Integrals . . . 1065
22.6 Quantum Field Theory . . . 1066
22.6.1 The Electromagnetic Force from Symmetry . . . 1078
22.7 Unification Theories . . . 1094
22.7.1 Kaluza-Klein Theories . . . 1094
22.7.2 Grand Unified Theories . . . 1095
22.7.3 Supersymmetry & Supergravity . . . 1096
22.7.4 String Theory . . . 1097
22.7.5 The Empirical Myth? . . . 1100
22.8 Our Amazing & Expanding Universe . . . 1102
22.8.1 The Robertson-Walker Metric & Inflation . . . 1106
22.9 Problems . . . 1111

A Projectile Motion with Air Resistance **1113**
A.1 Solution by First Approximation . . . 1115
A.2 Solution by Series Expansion . . . 1123
A.3 Solution by Numerical Integration . . . 1129

B Energy & Momentum Conservation **1135**

C Proofs of Kepler's Laws **1141**
C.1 The First Law . . . 1141
C.2 The Second Law . . . 1146
C.3 The Third Law . . . 1147

D The Linearity of Lorentz Transforms **1149**

E Proof That $1+2+3+4+\cdots=-\frac{1}{12}$ **1155**

F Gratuitous Pictures of Field Lines **1163**

Afterword **1171**

A Bibliography of Sorts **1173**
Books About Physics 1173
Real Physics Books 1174

Benediction **1177**

Index **1179**

Acknowledgements

This text was written in LaTeX 2$_\varepsilon$ in the Emacs editor on boxes running Linux and OpenBSD. Figures were done with Xfig and gnuplot. Some calculations were checked by the symbolic-manipulation program Maxima running on CLISP. Octave was used for numerical integrations. All of these applications are free and open-source software. No part of this text has been tainted by the proprietary software of evil monopolists.

Spiritual Canon

Though a physics text provides only very limited scope for exercises of the soul, we have tried to write this text in the spirit of Aristophanes, one of the very few able to transcend the egotism and petty delusions of the human species and see things for what they really are.

Spiritual Cannon

A physics text does, however, provide great scope for firing cats out of cannons, and we will be doing this sort of thing at every opportunity.

Part I

Preliminaries

Chapter –1

Preliminaries

> Yesterday, I awoke with the whole day ahead of me, so I rolled over and went back to sleep. I mean, who needs that kind of pressure?
>
> J.B. (otherwise unknown)

–1.1 About the Course

You are born. Then for a long period of time nothing makes any sense at all. Then you die. Try to look upon this course as a small but integral part of that experience.

–1.2 About This Book

The principal criteria by which any introductory physics textbook is judged are, of course, its logically disciplined approach to the subject; its full and easily understandable explanations, discussions, and step-by-step derivations of everything from fundamental principles; its profusion of lucid examples and clarifying illustrations; and its abundance of enlightening exercises, at all levels of difficulty, to strengthen and extend the student's grasp of the physics. As the reader will quickly discover, what distinguishes this book from the myriad others already available is its wantonly perverse disregard for all of these virtues. But then you can't have everything. We think the important thing is that we've had fun writing it.

Its subject matter being full as weighty as, and not dissimilar to, that of Lucretius's didactic epic *De Rerum Natura*, we had considered composing this entire volume in pentameter, along the lines of

No matter how two meeting masses swerve,
They take due care momentum to conserve.

and so on. The violent exertion of producing even this modest couplet has, however, left our flushed and sweaty Muse too winded and exhausted for us to further contemplate such a grand endeavor — a lack of stamina that may perhaps not occasion unendurable disappointment in the reader. Yet even as a prose work this book is not without artistic virtue. Indeed, if one were to judge from what one sees in museums of modern art, apparently *anything* qualifies as art these days. And so we suppose that we could, in this modern spirit, nonetheless still declare this humble tome to be a work of art. Though you could also consider it a labor of love, if you're willing to count love of evil.

But let us turn from questions of art and love to reality for a moment. This book is very much a work in progress; it is still in a very nascent and primitive state. But, hey, you try finding the time and energy to write all this stuff. And don't even get us started about how long it takes to do even simple diagrams. Woof. Anyway, we hope that in conjunction with what we do in class this book will nonetheless serve your needs well enough that you won't need any other text. At any rate, this is all you will have to fly on, so you'll have to make do.[1] If you find errors or have suggestions for improvements, by all means let us know.[2]

At the end of each chapter are two sections, one consisting of problems, followed by another consisting of unelaborated, bottom-line "answers" to selected problems, against which you can check your own results. Be aware, however, that

- Sometimes it is possible to get the correct numerical or algebraic answer to a problem by the wrong logic. If your answer agrees with the given one but you are not completely confident of your understanding of the problem, be sure to come for help or ask about it in class.

- These answers are still new enough that there are going to be typos and various other oversights. If your answer disagrees with a given answer, you may well be the one who's right.[3]

[1] Actually, if you feel the need for greater elucidation or a more detailed exposition on a particular topic, you can certainly check out one or more of the introductory physics texts in the library, where you should be able to find quite a number of texts written for courses that use calculus.

[2] This is, of course, a trivially easy matter if we have you in class, but we can also be contacted at `jbg52@columbia.edu`. If you do send an email, be sure that the word *PHYSICS* (all uppercase) appears somewhere in the subject line or your words of wisdom will irretrievably vanish into the spam filter.

[3] Developing confidence in your logic and reasoning is an important part of your scientific

- Two very important aspects of fully understanding the physics — aspects neglected by all the physics texts we have seen — are the abilities to recognize nonsense and to ascertain when a problem is insoluble. To help you develop these abilities, there are a few problems in the text that either cannot be solved or ask something that doesn't make sense. If you really understand the physics, you should be able to recognize these problems.

Many problems offer hints in footnotes. Ideally, you will have done your best with a problem before you resort to looking at those hints — the more you grapple with a problem on your own, the more you will improve your understanding.

And now, before we embark on our grand adventure, one final note. While we have, of course, done our best to do justice to the subject matter, we would be unconscionably remiss were we not to point out at the outset that those in search of truth and wisdom, not to mention wit and entertainment, would be far better off reading Henry Fielding's *Tom Jones*. We heartily recommend it.

–1.3 Good Karma

> Vanity of vanities, saith the Preacher, vanity of vanities; all is vanity. — *Ecclesiastes* 1:2.

How can I be successful in this course? you ask.

Do the reading. Sometimes the reading will make sense immediately, sometimes it will make sense only after we have talked about the physics it covers in class. Should you feel the need or the inclination, feel free to use other physics texts as well.

There are, of course, many demands on your time and energy. One consequence of this is that typically "The homework for tonight is to read pages . . . " is translated into "Woohoo! We don't have any homework tonight!" Neglect the reading at your own peril: to do well in the course, you need to understand the physics, and the point of the reading is to help you gain that understanding. We know that you have the best intentions in this regard, and to help you realize those good intentions, there may well be pop quizzes to check whether you've done the reading.

---

education. Contrary to certain fashionable political views, all points of view are not equally valid, and it is actually a virtue in science to be able to assert — politely, of course, and in the interest of furthering everyone's understanding — , "I'm right and you're wrong."

Do the homework problems. The homework problems are intended not only to give you practice, but also to make you *think* about the physics. It is only by thinking about the physics, by struggling with it, that you will develop an understanding and a working knowledge of the concepts. Some problems will be straightforward, others will be more challenging. On those occasions when you find yourself stuck on a problem, do not simply give up: make as much progress as you can with the part that you are stuck on and then make note (that is, actually write down) what it is that has you stuck. Then try to continue with the remaining parts of the problem.

In the past, we've often found ourselves in a vicious cycle: since the homework makes important points about the physics, we need to discuss it in class. The world being the place it is, people then often figure that they can blow off the homework — he's going to go over it in class, anyway, right? — which means we have to spend even more time on the homework in class, so that people are even less inclined to do it, and so on. To prevent this from happening, we will be regularly checking and sometimes even collecting the homework. You may even be called to the board to demonstrate your solutions to homework problems. And we will also be getting a very good idea of how well you are keeping up by observing you contributions to group work.

Always aim for understanding the concepts. This is not a course in plug-and-chug or monkey-see-monkey-do. To the extent possible, tests will ask you to demonstrate your understanding of the principles by asking you to explain them or to solve problems that require you to apply them in ways different from what you will have seen in homework and examples. If you really understand the physics, you should be able to reason your way successfully through these novel applications.[4] And your grades on the problems will correspondingly reflect our assessment of your understanding of and ability to apply the physics: to get credit, you need always to show your work or indicate your reasoning clearly. Arithmetic or algebraic errors are relatively minor; logic errors are relatively serious. You won't get much credit for pulling the right numerical answers out of the sky or arriving at them by incorrect reasoning.

[4]Most of the time, of course; no one is always going to get everything right, especially under time pressure. That's why tests are curved. (Curiously, it seems to have escaped the notice of many people that all tests are always curved: even if the test is graded on a traditional 100-point scale, someone has to decide what constitutes 100 points of work and how many points each part of the test is worth.)

Use consultation. With rare exception, we are in the classroom and available for consultation every consultation period that we do not have a lab.

When you come for help, we will try to help you think your own way through whatever is causing you difficulty, as opposed to simply telling you the answer. This process can sometimes be a bit painful (for both of us), but just telling you the answer isn't really going to help you understand the physics. And, historically, people who have come to consultation have realized improvement in their performance.

Our hope is that you will leave the course with a meaningful understanding of some physics. And even in the unlikely event that you never have occasion to make direct use any of the physics we cover, we hope that the course will have helped develop your general ability to reason logically and analytically. And with that, you can better figure out for yourself what is really meaningful in life and live accordingly.[5]

ὁ ἀνεξέταστος βίος οὐ βιωτὸς ἀνθρώπῳ.[6]

–1.4 The Zen of Problem Solving

Physicists don't express physical relations mathematically just out of spite; they do it because math is the natural language of physics.[7] If a picture is worth 10,000 words, then an equation is worth 10,000 pictures. Consider, for example, string theory: the action for the heterotic string can be written

$$S = -\frac{1}{2\pi}\int d^2\sigma\,\sqrt{-h}\Big(h^{\alpha\beta}\partial_\alpha X^\mu \partial_\beta X_\mu + i\overline{\psi}^\mu \rho^\alpha \partial_\alpha \psi_\mu\Big)$$

[5] This might seem self-evident, and yet it is shocking how little thought many people give to what they do with their lives. Too many blindly pursue the vainest and most superficial ambitions. They might as well be roaches or even stones — living by blind instinct, insensible of those things that truly are meaningful and worthwhile. Why so much importance placed on going to Harvard, Princeton, or Yale? On driving a Mercedes or a BMW? On becoming a doctor, lawyer, or CEO?

If you are worried about college admissions, maybe you have some thinking to do.

[6] "The unexamined life is not worth living," Plato's *Apology of Socrates*, 38a — which we would strongly recommend everyone read (though it's hard to find a translation that does justice to the original). And while we're at it, we'd also strongly recommend Voltaire's *Candide*.

[7] This is why all nonmathematical expositions written for people on the street, even those written by very capable and knowledgeable authors, do not convey any real understanding of physics; words are simply inadequate for the task. If you don't understand the math, you don't understand the physics.

If string theory is correct, this single equation accounts for *all* of the physics of our universe. Now try summing up all of the physics of our universe in words.

Understanding the physics boils down to understanding the mathematical relations by which it is expressed: you need to understand what physical quantity is represented by each of the variables and what each equation tells you about how those quantities are related. And the best way to develop this understanding is by solving problems.

Sometimes you will intuitively know how to do a problem as soon as you look at it. Sometimes, you may initially be more or less at a loss. In this latter case, you need to think your way through the problem logically and methodically, and there is a general procedure that will help guide your thinking:

Draw a picture. Often making a simple sketch will help you visualize what's going on.

List the quantities involved. Making a simple table of the variables involved and the physical quantities to which they correspond, and noting which are known and unknown, will help you see at a glance what you have to work with.

If you define any variables of your own, be sure to make your notation clear. If, for example, you use v_c to represent the velocity of a cat, then you should say something like "v_c = velocity of cat." Doing so is not only essential if your work is to be intelligible to someone else — such as, say, the person grading it — , but will also help clarify your own thinking.[8]

Work from the relevant general principles and relations. It may not be immediately clear to you precisely which principles and relations a solution requires, but the fundamental physical principles and the equations that express them are few in number, and listing those that may be relevant will often help you see which will enable you to make progress.

A common error (especially under the pressure of an exam) is to start wildly thrashing about, writing down more or less loosely connected

[8]The reader will here note that we habitually not only put a space on either side of a dash, but also retain any punctuation marks that would have been present without the dash. We do this because we believe it a Right and Good Thing. We also habitually punctuate lists as *A, B, and C* and see nothing wrong with splitting infinitives. Those who disagree with these usages are invited to inquire where they can go. We may, however, justly be faulted for irregularities in capitalization and a frustratingly incorrigible tendency to hyphenate what should properly be either a single or two separate words. Notice of particular instances of such errors is welcome.

heaps of numbers or algebraic expressions, hoping that the solution will fly out from them like partridges flushed out from the bushes. Don't do this. *Never start by writing down numbers or secondary expressions; always first write down, in terms of algebraic variables, the general relations relevant to the problem.* You will be far less likely to make mistakes, and far more likely to find the path to enlightenment, if you work from general principles than you will if you try to pull a solution out of the sky.[9]

Work in terms of variables. As we get further into the course, you will increasingly find that the problems don't even give you any numbers to work with; rather, you will have to keep track in your mind of which variables represent known quantities and which unknown quantities, and solving the problems will involve arriving at expressions for the latter in terms of the former. Developing the ability to solve problems algebraically rather than numerically is a large and critical step toward doing big-people physics. Even when you have numerical values for known quantities, it is highly desirable to solve problems algebraically and postpone the numerical substitutions until the very last step. In addition to even more compelling reasons that will be explained shortly, it is much easier to go back and fix mistakes when you have worked in terms of variables rather than numbers: working in terms of numbers is like having shag carpeting in the kitchen.

Keep your eye on the target. Another large and critical step toward doing big-people physics is learning how to make incremental progress even when you can't see all the way through to the ultimate solution. Having recognized the relevant general principles and written down the corresponding relations, you should try to chip away at the problem: what new expressions can you derive and what calculations can you do that will bring you closer to where you want to go?

Developing your ability to make such incremental progress, more or less in the dark but toward a clear goal, is by far the most important and meaningful thing we will do all year. It is as much an art as a skill. This is where the real Zen comes in. Fortunately, however, this ability can, like any art or skill, be cultivated through earnest practice. And there are even a few things you can do to make the process easier for yourself:

- Be rigorously logical and methodical. Be disciplined in your thinking. Do not write down expressions for quantities in the hope that

[9] Or whatever you are in the habit of pulling your solutions out of.

they might be true; work things out so that you *know* that they are true.

A common error is to engage in random algebra, aimlessly doing whatever calculations come to mind in the almost always vain hope that some good will come out of it. *Never do algebra unless you have a plan.* You can mess around all you want with three equations and four unknowns, but it won't get you anywhere.

Even when the goal is clear and definite, often there are easier and harder ways to do the math. Don't just start flailing away at the calculation with a stick; the time taken to think about the easiest, most efficient way to do the calculation will almost always more than pay for itself — even (in fact, especially) on tests.

- Structure your work on the page to reflect the logic. The same people who do random algebra also tend to splatter their work over the page like a plate of spaghetti thrown against a wall. Your work should never squirm and writhe around the page, nor should it pass through wormholes from one part of the page to another. The more logically structured your work, the less likely you will make mistakes and the more likely you will find the path of enlightenment.
- Do some struggling. You'll never develop your ability if you just throw your hands up and give up when you're at a loss. To really learn, you have to spend some time scratching your head over the difficult problems.

Stop and think. Don't go gleefully skipping off when you have arrived at a result for the unknown. Think about what that result means physically and how it behaves: Does it make sense? Is it acceptable to have gotten a negative answer? Are the dimensions and units correct? Reflecting on what your result means is critical to your understanding of the physics and often will also help you catch errors.

This is the principal reason why it's highly desirable to work in terms of variables rather then numbers: only when your result is in terms of variables can you ascertain how your solution will behave for certain values or limits of the parameters.

Just how you go about analyzing your result depends on the problem and will, we hope, be more clear after you read the example that follows. In general you want to ask yourself questions like, What happens when one of the masses is really large or really small? Or when two masses are equal? Or when an angle is 0 or $\frac{\pi}{2}$? To help you develop this ability to analyze your results, many of the problems will explicitly ask you what happens in various cases.

We cannot emphasize enough how beneficial adherence to the above precepts will be to developing both your understanding of the physics and your ability to apply it. And we say this not only because of the importance of these precepts, but because long experience has shown that no advice is more likely to fall on deaf ears. At times, teaching can be like reading Shakespeare to your dog. It's depressing, really. But you can make a difference: heed these precepts, and you will go a long way toward restoring our faith in humanity. Not to mention developing both your understanding of the physics and your ability to apply it.

–1.4.1 An Unfortunate Example

What we need now is a simple, easy example that doesn't involve any physics you don't yet know and that illustrates everything we just said about the Zen of problem solving. I wish I could think of a problem that satisfies all those constraints. I really do. But the only problem I can think of that doesn't involve any physics you don't yet know and that illustrates everything we just said about the Zen of problem solving is not quite as simple as one would like. While this is unfortunate, we trust that you will be able to keep your focus on the techniques that the example is intended to illustrate and not let its apparent complexity cause you to stampede to the Registrar's Office like panicked wildebeest.

And so: Suppose you want to get to a point directly across a conveniently rectilinear river of width ℓ, from point A to point B, as shown, rather melodramatically, in fig. (–1.1). You can row a boat at speed v_r, the river's current flows to the right at speed $v_c > v_r$ (that is, the current is faster than you can row), and you can walk along the shore of the river at speed v_w. We will try to determine the angle θ at which you should row (that is, "aim" the boat) in order to get from A to B in the shortest time. Because the current will be carrying you downstream as you row, we expect that the fastest way will not be to row straight toward B ($\theta = 0$) but rather to row somewhat

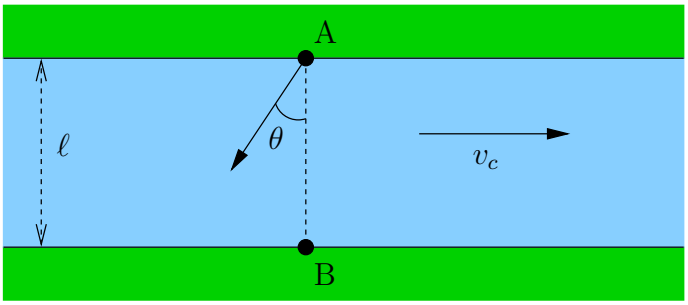


Figure –1.1: The River of Death

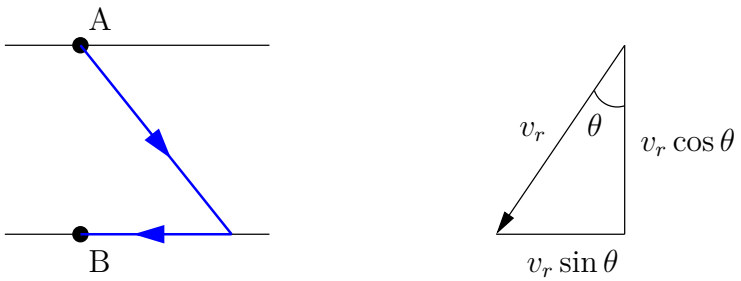


Figure –1.2: Your Actual Path and the Components of v_r

"into" the current, thereby partially canceling out its tendency to carry you downstream. Because $v_c > v_r$, this cancellation can only be partial; you will be carried somewhat to the right and thus have to walk back to point B along the river's shore, so that your actual path will be more or less as shown on the left side of fig. (–1.2).

The logic is thus: we are given three speeds (v_c, v_r, and v_w), and speed is distance over time:

$$\text{speed} = \frac{\text{distance}}{\text{time}}$$

Together with our knowledge of the distance ℓ across the river, this basic relation for speed should enable us to express the time to get from A to B in terms of the angle θ at which you row. Our first task is to work this relation out more precisely.

The time to get from A to B will be the sum of the time to reach the far shore and the time to walk along the shore from your landing point to point B. We know the distance perpendicularly across the river is ℓ, so if we can get a result for your speed in the direction perpendicularly across the river we can figure out the time it will take to reach the far shore. As you can see from the right side of fig. (–1.2), the component of your rowing speed that is carrying you perpendicularly across the river is $v_r \cos\theta$. The time t_{cross} that it will take you to reach the far side will therefore be

$$\text{time} = \frac{\text{distance}}{\text{speed}}$$

$$t_{cross} = \frac{\ell}{v_r \cos\theta}$$

If we knew how far you will have to walk along the shore, we could, knowing that the speed at which you can walk along the shore is v_w, figure out the time it will take you to get from your landing point to point B. And now that we know how long it will take to cross the river, we can figure out how far you will have to walk along the shore: from the right side of fig. (–1.2), you can see that the component of your rowing speed that is working against the current is $v_r \sin\theta$. This means that speed at which you are being carried

to the right is not the full v_c but rather $v_c - v_r \sin\theta$. During the time that it will take you to reach the other side, the distance d_{drift} that you will be carried downstream is therefore

$$\begin{aligned}
\text{distance} &= \text{speed} \times \text{time} \\
d_{drift} &= (v_c - v_r \sin\theta)\, t_{cross} \\
&= (v_c - v_r \sin\theta)\, \frac{\ell}{v_r \cos\theta} \\
&= \frac{v_c - v_r \sin\theta}{v_r \cos\theta}\, \ell
\end{aligned}$$

Walking at speed v_w, the corresponding time t_w that you will have to spend walking along the shore back to point B will thus be

$$\begin{aligned}
\text{time} &= \frac{\text{distance}}{\text{speed}} \\
t_w &= \frac{d_{drift}}{v_w} \\
&= \frac{\dfrac{v_c - v_r \sin\theta}{v_r \cos\theta}\, \ell}{v_w} \\
&= \frac{v_c - v_r \sin\theta}{v_w v_r \cos\theta}\, \ell
\end{aligned}$$

The total time t_{AB} to go from A to B will therefore be

$$\begin{aligned}
t_{AB} &= t_{cross} + t_w \\
&= \frac{\ell}{v_r \cos\theta} + \frac{v_c - v_r \sin\theta}{v_w v_r \cos\theta}\, \ell \\
&= \frac{\ell}{v_w v_r \cos\theta}(v_w + v_c - v_r \sin\theta)
\end{aligned}$$

where we have neatened up a bit by pulling out an overall factor of $\ell / v_w v_r \cos\theta$.

Our first task is now complete: we have an expression for the time to get from A to B in terms of θ and given quantities. Next we want to minimize this time t_{AB} with respect to θ, which is a straightforward if somewhat tedious maxima-minima problem:

$$\begin{aligned}
0 &= \frac{dt_{AB}}{d\theta} \\
&= \frac{d}{d\theta}\left(\frac{\ell}{v_w v_r \cos\theta}(v_w + v_c - v_r \sin\theta)\right) \\
&= \frac{\ell}{v_w v_r}\, \frac{d}{d\theta}\left(\frac{1}{\cos\theta}(v_w + v_c - v_r \sin\theta)\right)
\end{aligned}$$

If we divide both sides by the annoying factor $\ell/v_w v_r$ and do out the derivative, we have

$$\begin{aligned}
0 &= \frac{d}{d\theta}\Big(\frac{1}{\cos\theta}(v_w + v_c - v_r\sin\theta)\Big) \\
&= \left(\frac{d}{d\theta}\Big(\frac{1}{\cos\theta}\Big)\right)(v_w + v_c - v_r\sin\theta) + \frac{1}{\cos\theta}\frac{d}{d\theta}(v_w + v_c - v_r\sin\theta) \\
&= \frac{\sin\theta}{\cos^2\theta}(v_w + v_c - v_r\sin\theta) + \frac{1}{\cos\theta}(-v_r\cos\theta)
\end{aligned}$$

Solving this for θ might look awful, even hopeless, but a miracle occurs when you pull out an overall factor of $1/\cos^2\theta$ and simplify:[10]

$$\begin{aligned}
0 &= \frac{\sin\theta}{\cos^2\theta}(v_w + v_c - v_r\sin\theta) + \frac{1}{\cos\theta}(-v_r\cos\theta) \\
&= \frac{1}{\cos^2\theta}\Big((v_w + v_c - v_r\sin\theta)\sin\theta - v_r\cos^2\theta\Big) \\
&= \frac{1}{\cos^2\theta}\Big((v_w + v_c)\sin\theta - v_r(\sin^2\theta + \cos^2\theta)\Big) \\
&= \frac{1}{\cos^2\theta}((v_w + v_c)\sin\theta - v_r)
\end{aligned}$$

which yields

$$\sin\theta = \frac{v_r}{v_w + v_c} \tag{–1.1}$$

This is our result for the θ that will minimize the time to get from A to B.

As a quick check on our algebra, we can look at the dimensions in our result: since angles are dimensionless quantities,[11] the θ and hence $\sin\theta$ on the left side of eq. (–1.1) are dimensionless. In the denominator on the right side, we are adding two speeds, which is fine — if we instead had something like $v_w + \ell$, that would indicate an error somewhere earlier in the calculation, because it doesn't make any sense to add a length to a speed (which is length/time). And overall on the right side of eq. (–1.1) we have the ratio of two speeds, which, being dimensionless, matches the left side of eq. (–1.1).[12] So everything is okey-dokey dimension-wise. This does not prove that our

[10] Whether this constitutes proof of a divine being is left for the reader to decide.

[11] Degrees, radians, grads, etc., are not physical units; they're just reminders of how you're measuring your angles. Angles represent ratios of arc length to radius ($s = \theta r$) and are therefore dimensionless; properly they should always be measured in what we call radians, because $C = 2\pi r$ gives 2π as the angle corresponding to a full circle. But again, radians do not have dimensions: since the length of the circumference C and the radius r in $C = 2\pi r$ cancel out, the 2π radians in a full circle have no dimensions.

[12] It turns out that there are only three fundamental physical dimensions: length, mass, and time. The dimensions of all physical quantities are some combination of these basic three. When doing a dimensional analysis, it is customary to denote length, mass, and

result is correct, but had we found a mismatch in dimensions, that would definitely have indicated an error.

Now that we have a result in which we are fairly confident, we need to try to understand it. The first thing to note is that, of the givens, θ does not depend on the width ℓ of the river. This should seem sensible when you consider that, since ℓ is the only distance in the problem, and since

$$\text{time} = \frac{\text{distance}}{\text{speed}}$$

all times must be proportional to ℓ, which therefore constitutes merely an overall factor that divides out when we minimize the time. That is, changing the width of the river by a factor will scale all the distances and times, including the time to get from A to B, by that same factor, but it will not change the directions and angles of the path that minimizes t_{AB}.

What our result for θ does depend on are the speeds v_c, v_r, and v_w, in the form of the ratio

$$\frac{v_r}{v_w + v_c}$$

This suggests that it might be interesting to look at what happens when v_c, v_r, or v_w is very small or very large.[13] If $v_w \to \infty$, then $\sin\theta \to 0$ and hence $\theta \to 0$, which corresponds to rowing ("aiming") straight across the river. This what we would have expected: when you can walk really fast, it doesn't matter how far along the shore you have to walk, so you should simply set the course that will minimize the time to reach the other side, without worrying about how far you will drift because of the current.

When v_r is very small ($v_r \to 0$) or v_c very large ($v_c \to \infty$), we again get $\theta \to 0$: when your rowing speed is very small or the current is very swift, you aren't going to save yourself any time by trying to offset the current; you're better off just rowing ("aiming") directly across and then walking however far you have to along the shore to get back to point B.

We could look at other limits, but these are sufficient to illustrate how you go about interpreting and understanding your final result when solving problems.[14]

---

time by $[\ell]$, $[m]$, and $[t]$. Thus a speed would have dimensions $[\ell]/[t]$, and the dimensions of eq. (–1.1) would be

$$1 = \frac{\frac{[\ell]}{[t]}}{\frac{[\ell]}{[t]} + \frac{[\ell]}{[t]}} = 1 \quad \checkmark$$

[13] If the denominator were $v_w - v_c$ instead of $v_w + v_c$, the case $v_w = v_c$ would also be interesting to consider.

[14] For those who are curious, we cannot look at the limit $v_w \to 0$ without reworking the problem: although this limit seems to yield $\sin\theta \to v_r/v_c$, on our way to our result for $\sin\theta$ we divided by v_w. We also cannot look at $v_c \to 0$ or $v_r \to \infty$ without reworking

If you go back through the Zen principles of problem solving, you will see that the above solution adheres to them all: we drew pictures to help visualize the situation, made note of known and unknown quantities, and clearly defined our notation. We used the definition of speed and a little trigonometry to obtain a result for the time to get from A to B in terms of the known quantities, then minimized this time with respect to θ by doing a maxima-minima calculation. We also looked at the quantities on which our result for θ depended and how that result behaved in various limits of the given velocities and saw that everything made sense.

Also note that we used notation that reflected the kinds of quantities represented: t_{cross} for the *time* to cross, d_{drift} for the *distance* you drift downstream, etc. If you are in the habit of labeling every unknown quantity x, you should break that habit.

–1.5 Significant Figures

Now, no one really enjoys significant figures, but then that really isn't the point, is it?

No measured quantity is ever known with perfect accuracy, and significant figures allow us, at least in some primitive, rudimentary way, to take this into account. Suppose, for example, that you were so bored that you were reduced to trying to determine the area of a piece of xerox paper by measuring its length and width. If your ruler allows you to measure down to a millimeter and indicates that the length and width are 27.9 cm and 21.6 cm, for the area your calculator will give you $27.9 \times 21.6 = 602.64\,\text{cm}^2$. But quoting the five-figure result $602.64\,\text{cm}^2$ would imply that you were able to determine the area to an accuracy of about 1 part in 100,000, when in fact this result was derived from measurements each of only three figures and therefore accurate to only about 1 part in 1000. Since your result for the area is only as accurate as the data that went into calculating it, in this case you should quote an area accurate only to 1 part in a 1000 — that is, you should round off your result for the area to three figures: $603\,\text{cm}^2$.

---

the problem because we set everything up on the assumption that $v_c > v_r$. If we were to rework things for the case $v_r \sin\theta > v_c$, our speed parallel the river would instead be $v_r \sin\theta - v_c$, and our solution for θ would become

$$\sin\theta = \frac{v_r}{v_c - v_w}$$

This would seem to yield $\sin\theta \to -v_r/v_w$ when $v_c \to 0$ and $\sin\theta \to \pm\infty$ when $v_r \to \infty$, but we would have to remember that in maxima-minima calculations it is possible that the maximum or minimum is at the endpoints of the domain of the parameter: in this case θ is restricted to $0 \le \theta < \frac{\pi}{2}$, and the minimum in these two limits would actually be at $\theta = 0$, as expected.

The rules and uses of significant figures should already be familiar to you from previous courses in science; here we will merely refresh your memory a bit:

- For all mathematical operations *except* addition and subtraction, the number of significant figures in a *final* result is the number in the *least* accurately known data. (The annoying case of addition and subtraction is dealt with below.)

  Thus $1.23 \times 4.5 = 5.5$ as a final result.

- Keep one extra figure in *intermediate* results, to avoid rounding error.[15]

  Thus $1.23 \times 4.5 = 5.54$ as an intermediate result that you are going to use in a subsequent calculation.

- Zeros on the left ("leading zeros") don't count toward significant figures; zeros on the right ("trailing zeros") do.

  Thus 0.00123 has three significant figures: a length of 1.23 mm is accurate to 1 part in 1000 and does not become more or less accurate simply by being expressed in other units, such as 0.00123 m. On the other hand, 0.4500 has four significant figures: 0.45 would mean a value of $\frac{45}{100}$ to an accuracy of 1 part in 100, while 0.4500 means a value of $\frac{45}{100}$ to an accuracy of 1 part in 10,000.

- Pure numbers (like π or $\sqrt{2}$) are exact and do not affect the number of significant figures in the calculation.

  Alternatively, you can think of pure numbers as having an infinite number of significant figures. In practice, you simply express them to as many digits as necessary to prevent them from limiting the significant figures in your calculation. For example, in a calculation with four significant figures, it is enough to use 3.1416 for π.

- Don't count the digits in the power of ten in scientific notation, for example, the 10^{-6} in 4.32×10^{-6}; these powers of ten are equivalent to leading zeros.

- If a number lacks a decimal point, there are passionately differing opinions about the significance of trailing zeros: some would argue vehemently that 500 has three significant figures, others equally vehemently that it has only one. All of this just goes to show how silly people can

[15] Humanity, which distinguishes itself from the lower animals chiefly by its capacity for irrationality, seems to persist in the belief that the rounding used with significant figures *causes* rounding error. Be sure to remember (especially when doing lab work) that the rules for significant figures do not cause rounding error; they *prevent* it.

sometimes be. Our convention will be to be reasonable about it and show the decimal point when it matters. And of course you can always avoid this ambiguity by using scientific notation: the number of significant figures in 5×10^2, 5.0×10^2, and 5.00×10^2 are clear.

Significant figures are determined differently in the case of addition and subtraction.

One complication is that a subtraction may wipe out some of what would otherwise have been significant figures by making them into leading zeros. Suppose, for example, that you were trying to determine the length of an object and found that its 'start' and 'end' were at 12.6 cm and 11.8 cm along a meter-stick. Since each of these measurements has three significant figures, you might be tempted to give $12.6 - 11.8 = 0.800$ cm for the object's length. This would, however, imply that you had determined to object's length down to 0.001 cm, when in fact your measurements were only accurate to 0.1 cm. The problem is that the subtraction has made leading zeros out of the first two digits in the 12.6 and 11.8. So your result for the object's length really has only one surviving significant figure and should be quoted as 0.8 cm.

If only that were the sole problem presented by addition and subtraction. But, alas, there is a second complication. Suppose, for example, that you are taking the difference between 2.45 cm and 0.011 cm. Your calculator will give you $2.45 - 0.011 = 2.439$. Since the 2.45 has three figures and the 0.011 two, you might reckon that the difference has two figures and should be quoted as 2.4 cm. The problem with this reasoning is that the 0.011 is so much smaller than the 2.45 that subtracting it affects the 2.45 only starting at the third figure, so that all three of the original significant figures in the 2.45 remain significant. The difference should therefore be quoted as 2.44 cm. The same sort of argument would lead you to quote 2.46 for the result of $2.45 + 0.011$.

Both kinds of complications can be taken into account by means of the following (admittedly rather opaquely phrased) prescription:

- When numbers are added or subtracted, the position of the leftmost of the last significant figures of these numbers is the position of the last significant figure in the result.

Say what? you say. Some examples:

23.6	23.6	2.3	23.6
$-$ 21.4	$-$ 22.8	$-$ 0.284	$-$ 3.88
2.2	0.8	2.0	19.7

And when exponents are involved, just express the numbers in terms of a common power of ten:

$$2.34 \times 10^8 - 3.6 \times 10^6 = (2.34 - 0.036) \times 10^8 = 2.30 \times 10^8$$

So much for significant figures.

–1.6 Units & Conversions

Don't forget to indicate the units on any data or result that has physical dimensions. For example, simply reporting "4" for the height of an object is meaningless: 4 what? Inches? Meters? Miles?

We will usually stick with the MKS (SI) system of units, so that all of the physical quantities we deal with will be measured in meters, kilograms, seconds, or some combination of these basic three units.[16]

In conversions, do multiplications and divisions of units just as you do ordinary arithmetic with numbers. For example, to convert 18 inches into feet, do something like

$$18 \text{ in} \times \frac{1 \text{ ft}}{12 \text{ in}} = 1.5 \text{ ft}$$

Likewise, to convert cubic feet to cubic inches,

$$1 \text{ ft}^3 \times \left(\frac{12 \text{ in}}{\text{ft}}\right)^3 = 1728 \text{ in}^3$$

Doing conversions in this kind of explicit detail might feel really Mickey-Mouse, but it will save you from the dreaded and all too common error of getting the conversion factors upside down.

Conversion factors can usually be found in the jackets or appendices of textbooks; in the next section you will find the values of some of the more common ones.

–1.7 Conversion Factors, Prefixes, & Physical Constants[17]

Length

1 km	=	0.6214 mi		
1 m	=	3.281 ft	=	39.37 in
1 in	=	2.540 cm		
1 mi	=	5280 ft	=	1.609 km
1 ly	=	9.461×10^{15} m		

[16] MKS stands for (*doh!*) meter-kilogram-second, and SI for the French phrase for International System.

[17] Values of the physical constants and astronomical values are mostly from the various documents available through Lawrence Berkeley National Laboratory (`http://pdg.lbl.gov/2005/reviews/contents_sports.html`). Conversion factors are from the freely redistributable open-source Units program (`http://www.gnu.org/software/units/units.html`), with some rounding. A few odd values we just made up.

Volume

$$1\,\text{gallon} = 3.785\,\text{liters}$$
$$1\,\text{liter} = 1000\,\text{cm}^3 = \tfrac{1}{1000}\,\text{m}^3$$

Speed

$$1\,\text{mph} = 0.4470\,\text{m/s}$$
$$60\,\text{mph} = 88\,\text{ft/sec}$$

Force

$$1\,\text{N} = 0.2248\,\text{lb}$$
$$1\,\text{lb} = 4.448\,\text{N}$$

Pressure

$$1\,\text{Pa} = 1.450\times10^{-4}\,\text{lb/in}^2$$
$$1\,\text{atm} = 1.01325\times10^{5}\,\text{Pa} = 760\,\text{mm Hg} = 14.70\,\text{lb/in}^2$$
$$1\,\text{mm Hg} = 133.3\,\text{Pa}$$

Energy & Power

$$1\,\text{cal} = 4.1868\,\text{J}$$
$$1\,\text{BTU} = 1055\,\text{J}$$
$$1\,\text{eV} = 1.60217653\times10^{-19}\,\text{J}$$
$$1\,\text{hp} = 745.7\,\text{W}$$

Prefix	Factor	Prefix	Factor
		c (centi)	10^{-2}
k (kilo)	10^{3}	m (milli)	10^{-3}
M (mega)	10^{6}	μ (micro)	10^{-6}
G (giga)	10^{9}	n (nano)	10^{-9}
T (tera)	10^{12}	p (pico)	10^{-12}
P (peta)	10^{15}	f (femto)	10^{-15}
E (exa)	10^{18}	a (atto)	10^{-18}

Body	Mass (kg)	Radius (m)	Mean Orbital Radius (m)	Period (days)
Sun	1.98844×10^{30}	6.961×10^{8}		
Earth	5.9723×10^{24}	6.378140×10^{6}	$1.49597870660 \times 10^{11}$	365.24219
Mars	6.4185×10^{23}	3.3774×10^{6}	2.279×10^{11}	686.9600
Moon	7.347673×10^{22}	1.7360×10^{6}	3.84403×10^{8}	27.32166155

Physical Constants

Speed of light	c	=	2.99792458×10^{8} m/s
Electron charge	e	=	$1.60217653 \times 10^{-19}$ C
Universal gravitational constant	G	=	6.6742×10^{-11} Nm^2/kg^2
Planck's constant	h	=	$6.6260693 \times 10^{-34}$ Js
	$\hbar = h/2\pi$	=	$1.05457168 \times 10^{-34}$ Js
Electron mass	m_e	=	$9.1093826 \times 10^{-31}$ kg
Proton mass	m_p	=	$1.67262171 \times 10^{-27}$ kg
Neutron mass	m_n	=	$1.67492728 \times 10^{-27}$ kg
Electric force constant	k	=	$8.987551788 \times 10^{9}$ Nm^2/C^2
Electric permittivity of vacuum	ϵ_0	=	$8.854187817 \times 10^{-12}$ C^2/Nm^2
Magnetic permeability of vacuum	μ_0	=	$4\pi \times 10^{-7}$ N/A^2
Boltzmann constant	k	=	$1.3806505 \times 10^{-23}$ J/K
Avogadro's number	N_0	=	6.0221415×10^{23}
Gas constant	R	=	8.3144727 J/mol·K
Atomic mass unit (u)			$1.66053886 \times 10^{-27}$ kg
Absolute zero			−273.15 °C
Acceleration due to gravity	g	=	9.80665 m/s^2

–1.8 Order-of-Magnitude Estimates

–1.8.1 An Example

It is possible to make rough but meaningful estimates of all sorts of quantities — things you might at first think you were clueless about — with just common knowledge and common sense. All of these order-of-magnitude estimates[18] are pretty much the same, so the procedure is most easily illustrated by example:

Suppose (to take a decidedly lame but classic example) you want to estimate the number of jelly beans in a jar — maybe a Kewpie doll is at stake, or tickets to the Final Four. You could simply directly estimate the number of jelly beans in the jar, but that would be just a wild guess; there's no way to get a good feel for such a large number: a hundred thousand? A million? A zillion? Who could say? To make a reliable estimate, you need to break the problem down into smaller pieces, small enough that you can easily visualize or get a feel for them.

If you knew the volume of the interior of the jar and the average volume taken up by a jelly bean, all you'd have to do is divide the former by the

[18] We use this coloring for technical terms when they appear for the first time.

latter and you'd have the number of jelly beans. So you need to come up with estimates for these two volumes.

First, the jelly bean: You could try to estimate the volume taken up by a single jelly bean directly, but for most people this would still be rather hard to get a good feel for: a cubic centimeter? A half an ounce? A tenth of a milliliter? You need to break the problem down still further: to derive a reliable result for the volume of a single jelly bean, you need to estimate its geometry and its corresponding dimensions.

Now, a jelly bean is actually a kind of balloon shape, but that's a complicated shape to deal with. You could approximate that the jelly bean is a little cylinder, in which case you'll need to estimate its diameter and length in order to get a result for its volume. Or you could approximate that the jelly bean is a little rectangular box, in which case you'll need to estimate a length, a width, and a height. Which choice is best? And in any case, what about the air space between beans? This is where the *order-of-magnitude* part of *order-of-magnitude estimate* comes in: order of magnitude refers to the power of ten (like the 4 in 10^4); in an order-of-magnitude estimate, you're only trying to get within a factor of ten of the actual answer. In the present example, that means it doesn't matter whether you treat the jelly bean as a balloon shape, a cylinder, or a box, or even whether you account for the (relatively) small air space between beans — the differences between the resulting volume estimates are not going to be significant. So you might as well choose the simplest shape to deal with: the box.

If the jar of jelly beans is right in front of you, you can just look at the beans when estimating their dimensions. But even if you don't have any beans to look at when you're working through the estimate, you can still visualize one in your mind. Let's suppose that, one way or the other, your estimate of the dimensions is $\frac{1}{3}\,\mathrm{cm} \times \frac{1}{3}\,\mathrm{cm} \times 1\,\mathrm{cm} = \frac{1}{9}\,\mathrm{cm}^3 \approx 0.1\,\mathrm{cm}^3$. Someone else might have somewhat different estimates for the length, width, and height and instead get $0.4\,\mathrm{cm}^3$. Another person might get $0.08\,\mathrm{cm}^3$. All of these estimates would be correct in the sense that they are based on reasonable estimates of the dimensions of a jelly bean; there is no one right answer for such a rough estimate, and you wouldn't even consider two estimates to differ significantly unless that difference was by a factor of ten or more. For this reason also we rounded off our result to one significant figure — which is really all we can justify in such a rough estimate.[19]

Even if the jar were opaque and you had no prior experience with jelly beans, you could still make an estimate of the volume of a jelly bean based only on the knowledge that jelly beans are eaten by the handful. Remember-

[19]There is, of course, such a thing as a bad estimate. To say that a jelly bean is like a cube 1 inch on a side would be a bad estimate because 1 inch × 1 inch × 1 inch is clearly unreasonably large.

ing that you are only trying to get within a factor of ten in your estimates, you could reason as follows: If we assume, for lack of any better information, that the geometry of a jelly bean is cubic, that won't be too far off even if it is in fact spherical or oblong or some other reasonable geometry. And each side of this box would reasonably be on the order of 1 cm: 10 cm on a side would clearly be too large (you'd have to take bites out of one); 0.1 cm = 1 mm on a side seems too small (it'd be like eating poppy seeds). Thus you would arrive at an estimate of 1 cm × 1 cm × 1 cm = 1 cm^3 for the volume of each jelly bean — not very different from the estimates obtained by someone who had actually seen jelly beans.[20]

Now for the jar: If you knew that the jar was a one-gallon jar, it would be both accurate and correct to use 1 gal = 3785 cm^3 for its volume. Otherwise, you have to do for the jar what you did for the jelly bean: estimate its geometry and dimensions. Suppose the jar looks cylindrical and that you estimate it has a diameter of 15 cm and a height of $\frac{1}{3}$ m $\approx$ 30 cm. Your estimate for its volume would then be $\pi r^2 h \approx 3 \times (15)^2 \times 30 \approx 2 \times 10^4$ cm^3. Your result for the number of jelly beans in the jar would then be

$$\frac{2 \times 10^4 \ \frac{\text{cm}^3}{\text{jar}}}{0.1 \ \frac{\text{cm}^3}{\text{bean}}} = 2 \times 10^5 \ \frac{\text{beans}}{\text{jar}}$$

Bear in mind that this is, of course, only a very rough estimate. The actual number of beans could easily be anywhere within about a factor of ten of this: 80,000, or 400,000, or even 345,262.

–1.8.2 General Points

- The key to making good order-of-magnitude estimates is breaking the problem down into pieces small enough that you can easily visualize or get a feel for them.
- When it is beyond your experience to give a knowledgeable estimate of a quantity, you can fall back on the "1, 10, 100" reasoning. (See footnote 20.)
- The point of the homework involving estimates is to give you practice making estimates. (*Duh!*) You may be able to find some of the results you need in books or, of all places, on the Internet, but that would be defeating the purpose of the homework — so don't do it.

[20] Similarly, suppose you needed an estimate of how long a walrus lives. You don't need to have any specialized knowledge of walruses: for a large mammal, one year is clearly too short, and for any mammal 100 years would be a very long time, so anything in between would, for lack of any better information, be reasonable: 10 years, 30 years, 45 years, whatever — these numbers are all of the same order of magnitude.

–1.8.3 A Brief Discourse on Malarkey

To some of you this estimate business may seem like malarkey.[21] Believe it or not there are, however, uses for it in physics and engineering: a quick order-of-magnitude estimate can often save you a great deal of design time by making immediately clear whether a project is feasible or not. From the jelly-bean estimate, for example, you would immediately conclude that it would not really be feasible to count the exact number of jelly beans in the jar by hand — 200,000 is just too many to count. You would also be skeptical of anyone who claimed to have counted all the beans by hand. More generally, if you are able to make your own order-of-magnitude estimates, you will often be able to test the validity of assertions made in news articles by reporters and politicians who think you're the most gullible animal in the barnyard. And still more generally, making estimates is good practice in basic problem solving — and nothing is more certain than that life will require you to do a great deal of problem solving.

[21] Then again, to some of you, this whole course probably seems like total malarkey. To the extent that this is true, however, the course is excellent preparation for life.

–1.9 Problems

1. Be obsessively careful of significant figures in all of the following.

   (a) Determine the circumference of a circle of radius 3.5 cm.

   (b) Determine the volume of a rectangular box that is 4.6 cm by 12.3 cm by 5.70 cm.

   (c) Determine the total mass of three objects whose individual masses are 1.23×10^2 kg, 3.4 kg, and 9.87×10^{-1} kg.

   (d) Three points A, B, and C lie along a straight line. Determine the distance between points A and B if the distance between points A and C is 14.7 m and the distance between points B and C is 3.62 m. (Dude: that means take the difference.)

2. The density of water is 1.0 g/cm^3. What is it in

   (a) kg/m^3?

   (b) lb/ft^3?

   Note: a pound is actually a unit of force (weight), not mass, but there is a correspondence 1 kg $\leftrightarrow$ 2.20 lb. (We'll discuss this in more detail when we get to forces.)

3. (a) Determine your age in *seconds*.

   (b) How exact is this result? That is, by how much could your result for the number of seconds be off? You should take into account how accurately you know when you were born (which certainly isn't likely be down to the precise second) and also how accurately you dealt with leap years.

   (c) How many significant figures does your result for your age in seconds therefore have?

4. Estimate the percent error in measuring

   (a) a mass of 20 g on a scale that gives a reading out to a hundredth of a gram.

   (b) a length of 20 cm with an ordinary meter stick (which, you will recall, has gradations of millimeters).

5. Suppose that for some silly reason you wanted to determine the volume of a circular cylinder of diameter 1.23×10^2 mm and height 0.025 m. What do you suppose that volume would be?

6. People who apparently are in a position to know tell us that the mass and radius of the Earth are 5.9723×10^{24} kg and 6.378140×10^3 km, respectively. Determine the Earth's average density in g/cm^3, explicitly noting any assumptions you are making.

Note that for problems #7 through 16, the idea is for *you* to make the estimate, not to test your skill at doing Internet searches or posing questions to your phone. You should be working solely from your common knowledge, without looking up the values of any quantities.

7. How much cow flop does a typical cow flop over its lifetime?

8. (a) How many drops of water are there in all the oceans combined?
   (b) While you're at it, estimate the number of water (H_2O) molecules currently in your body that were in the body of Socrates at the moment he died, assuming that the water molecules that were in Socrates are now evenly mingled among all the water molecules on Earth. (If you want, you can use Confucius or Cleopatra or whatever other ancient personage you regard as awesomely righteous.)

9. How many golf balls would be needed to completely fill the average Pizza Hut?

10. How much toothpaste will you use over the course of your lifetime?

11. How much water does the average family use per week?

12. One year it was proposed that the Bowl[22] be filled with Jello as the senior prank. How much would it cost (in terms of dollars spent for the Jello at the supermarket) to do this?

13. There is an incipient annual tradition in Malmedy, Belgium of making a 10,000-egg omelet in 4 m pan. How thick is the omelet, assuming that it is a rather boring omelet in which eggs are the only ingredient?

14. Estimate the number of toilets needed for a building with 1200 nine-to-five workers. (This sort of thing is what engineers really do for a living.)

15. Estimate the number of people in the world who are, at any given moment, eating (or engaged in any other lone or shared common activity that you find interesting).

16. (a) Make up (and perform) your own interesting, bizarre, or even macabre order-of-magnitude estimate.
   (b) Does the problem you came up with indicate a need for counseling?

[22] While we generally try to avoid references to a particular school, that simply wasn't possible in this case. But you can easily substitute some structure at your own school for the "Bowl."

–1.10 Sketchy Answers

(1a) 22 cm.

(1b) 3.2×10^2 cm^3.

(1c) 127 kg.

(1d) 11.1 m.

(2a) 1.0×10^3 kg/m^3.

(2b) 62 lb/ft^3.

(4a) 0.05%.

(5) 3.0×10^2 cm^3.

(6) 5.4950 g/cm^3.

Chapter 0

Optics

0.1 Light Waves

Light is actually a traveling electromagnetic wave, a rather abstract oscillation of electric and magnetic fields, but in many ways very similar to the ocean waves you see at the beach.[1] Both are examples of transverse traveling waves — traveling meaning simply that the wave moves forward (propagates), and transverse meaning that the wave oscillation is perpendicular to this direction of propagation. In the case of an ocean wave, the wave propagation is horizontally toward the shore, and the wave oscillation is the vertical bobbing up and down. If you do out the math (which isn't worth getting involved in at this point), the cross section of these waves turns out to be sinusoidal.[2]

Some basic parameters of waves:

Wavelength The perpendicular distance from one wave crest to the next is called the wavelength (symbol λ). (See fig. (0.1).)

Amplitude If you hold a meter stick vertically at a fixed location in the water, you will see the water rhythmically rise and fall, always reaching the same highest point at each wave crest and the same lowest point at each wave trough. Half this vertical distance (in other words, the distance from the *midpoint* to crest or to trough) is called the amplitude

[1] `http://www.phy.ntnu.edu.tw/java/emWave/emWave.html` has a decent animation of an electromagnetic wave.

[2] At least in this simple case of plane waves whose crests form straight lines. When you toss a pebble into a pond, you may have noticed that the "rings" of the wave emanating from the splash diminish in height and become closer together as they propagate outward. Their cross section is actually a Bessel function. And in three-dimensional cases like sound waves emanating from a point, the cross section is a spherical Bessel function. Messier cases have correspondingly messier functional forms. Even our simple planar ocean wave gets weird when it encounters an obstacle or enters shallow water. Nobody said life was simple.

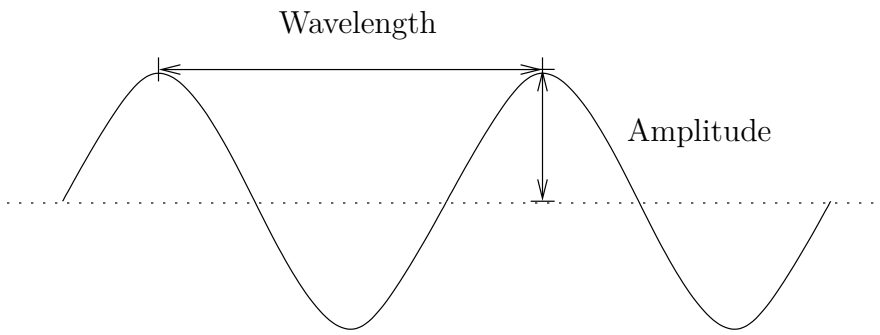


Figure 0.1: Cross Section of a Plane Wave

(symbol A). The amplitude gives a measure of how "big" the wave is. The midpoint of the oscillation is the level of still water, that is, where the surface of the water would lie if there were no wave. (See fig. (0.1).)

Period & Frequency The time it takes for the water level on the meter stick to complete one full cycle (say, to go from crest to trough and back to crest again) is called the period of the wave (symbol T). The reciprocal of the period is the frequency of the wave (symbol f): $f = 1/T$. (Don't see fig. (0.1).)

Since the time between successive wave crests is one period, and since the corresponding crest-to-crest distance covered by the wave as it travels forward is one wavelength, the speed of the wave's forward motion (its speed of propagation) is simply

$$v = \frac{\text{distance covered}}{\text{corresponding time}} = \frac{\text{wavelength}}{\text{period}} = \frac{\lambda}{T}$$

Since by definition the frequency $f = 1/T$, we can also write this relation for the speed as $v = \lambda f$. (In the context of light, the symbol c is conventionally used for the speed of light, and we write $c = \lambda f$.)

The period T is defined as the time to complete one full cycle. In MKS units, it would therefore be measured in seconds, or, if you will, sec/cycle.[3] The frequency f, being the reciprocal of the period, would therefore be in cycles/sec, sometimes abbreviated cps, and alternatively denoted by Hz (Hertz, $1\,\text{Hz} = 1\,\text{cps}$). The frequency is therefore a kind of rate and tells us how rapidly the wave oscillation is occurring.

The electromagnetic spectrum is continuous and unbounded — that is, the wavelength of light can have any value, no matter how big or small. You need not and should not memorize table (1), but it will give you a rough idea

[3] A cycle, like a degree or a radian, is not a physical unit; it's just a reminder of what sort of quantity we're talking about.

Part of Spectrum	Wavelength
Gamma Rays	$\lesssim 10^{-3}$ nm
X Rays	$\sim 10^{-2}$ nm to ~ 10 nm
Ultraviolet (UV)	~ 10 nm to 380 nm
Visible	violet 380–425 nm blue 425–520 nm green 520–565 nm yellow 565–590 nm orange 590–625 nm red 625–750 nm
Infrared (IR)	750 nm to ~ 1 mm
Microwaves	~ 1 mm to ~ 10 cm
TV & Radio	~ 10 cm to 1 km or more

Table 1: The Electromagnetic Spectrum

of parts of the spectrum and their relative wavelengths. We say rough idea not only because finer divisions than those shown here have been defined for the spectrum — there are, for example, hard and soft X rays, and oodles of different subranges of TV and radio waves — , but also because different authors often give slightly different values for the upper and lower limits of the ranges of the various parts of the spectrum.

More interestingly, for the visible spectrum the correspondence between wavelength and color actually varies slightly from one person to another: one person's pale green may be another's robin's-egg blue, etc.[4] Without getting too far into the physiology,[5] the light sensors on your retina are of two types: rods and cones. The rods are more sensitive than the cones, but can distinguish only light versus dark and not color. This is why in dim light, when your vision is relying mostly on the rods, you can see only shades of gray. Your color perception is due to the cones, which are of three subtypes — red, green, and blue — , each of which is stimulated most strongly by light of wavelengths close to its own "natural" wavelength (about 580 nm for red,[6] 540 nm for green, and 440 nm for blue). The color you perceive depends on the proportions in which the light incident on your retina excites each of these three types of cones, and this is the reason why it is possible to reproduce all

[4]This variation has nothing to do with color blindness, which is a totally separate issue; we are talking here about variations between individuals with normal color vision.

[5]Those interested in reading more about color vision might find `http://en.wikipedia.org/wiki/Color_vision` a good place to start.

[6]While, as you can see from table (1), a wavelength of 580 nm does not itself correspond to the color red, these cones do respond most strongly to wavelengths toward the red end of the visible spectrum; they are responsible for our perception of red because red light will stimulate these "red" cones, not the green or blue cones.

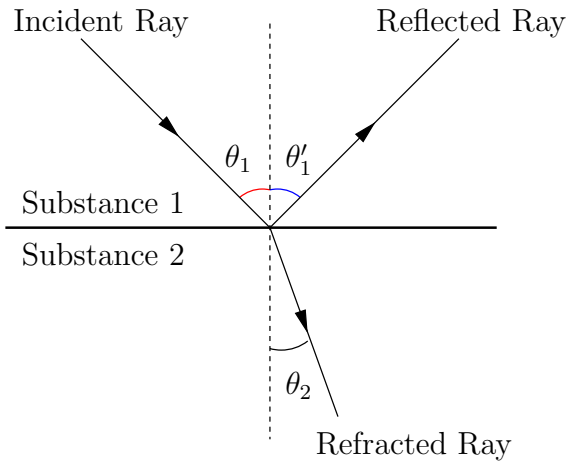


Figure 0.2: Incident, Reflected, And Refracted Rays

the colors perceptible to humans by mixing together just red, green, and blue light, as is done on television screens and computer monitors. Presumably we all have cones of the same chemistry and thus frequency sensitivity, but there are slight variations from one person to another in the proportions of cones in the retina and possibly in the neurological processing of the signals from the cones, which would lead to slight variations in color perceptions.[7]

0.2 Geometrical Optics

If you keep your focus on a particular point on a wave as it moves toward the shore — say, a point on a particular wave crest — , this point traces out a straight line, and the progress of that point on the wave can therefore be represented by an arrow or ray. The study of light by the tracing of such rays constitutes geometrical optics.

0.2.1 Reflection & Refraction

When light strikes an interface between two substances, in general the light is partly reflected back into the first substance and partly transmitted into the second substance. The transmitted light is said to be refracted because its direction of propagation generally changes as it passes from the first into the second substance, as shown in fig. (0.2).[8]

[7] Or so we surmise — our expertise is in physics, not physiology, and since we're too pressed for time to research the question, we're just making a reasonable guess. If you know better, please clue us in.

[8] A good interactive illustration of reflection and refraction can be found at `http://www.phy.ntnu.edu.tw/java/light/flashLight.html`.

The direction of propagation of the incident, reflected, and refracted rays are conventionally specified by giving the angles those rays make with the normal (that is, the perpendicular) to the interface. Thus in fig. (0.2) θ_1 is the angle of incidence, θ_1' the angle of reflection, and θ_2 the angle of refraction.

Light is an electromagnetic wave and therefore obeys the same relations as all other electromagnetic phenomena. It turns out (and we'll get to some of this later in the year when we study electromagnetism proper) that all of electromagnetism is governed by just four equations known as the Maxwell equations, which were conjured up in the 1860s by the brilliant Scotsman [9] James Clerk Maxwell. Using the Maxwell equations, it is possible to calculate the properties of the light that is reflected and refracted in terms of the electric and magnetic properties of the two substances, and in fact this is the proper, rigorous way to derive the relations governing reflection and refraction. Since the amount of physics and math required to take this approach is beyond what we can or would want to get into at this juncture, we will simply accept that the following turns out to be true:

For the reflected ray, the angle of reflection equals the angle of incidence ($\theta_1' = \theta_1$ in fig. (0.2)). This is known as the law of reflection.

For the refracted ray, things are more complicated. Among the electromagnetic properties of a substance is a (more or less) constant parameter governing the passage of light. This parameter, called the index of refraction and conventionally denoted by the symbol n, can be calculated from what are known as the electric permittivity and magnetic permeability of the substance, but we will treat it simply as a given quantity. For a vacuum, the index of refraction $n = 1$ exactly; for all other substances, $n > 1$ (though for rarified substances like air, $n \approx 1$). It turns out that the angle of refraction is related to the angle of incidence by

$$n_1 \sin\theta_1 = n_2 \sin\theta_2,$$

where, as shown in fig. (0.2), θ_1 is the angle of incidence in substance 1, θ_2 is the angle of refraction (transmission) into substance 2, and n_1 and n_2 are the indices of refraction of substances 1 and 2, respectively. This relation is known as the law of refraction or Snell's law. Refraction is responsible for the bending and similar distortions of perspective that you see when you look at a spoon partially submerged in a glass of water, look at things through bottles, or look into fish tanks from an angle.

Note that the law of refraction is symmetric under the interchange of the indices 1 and 2: if a light ray traveling at angle θ_1 in substance 1 emerges at angle θ_2 in substance 2, then a light ray traveling at angle θ_2 in substance 2 will emerge at angle θ_1 in substance 1. That is, the path followed by the light ray is reversible. If, for example, you are standing by the edge of a

[9] *Freedom!*

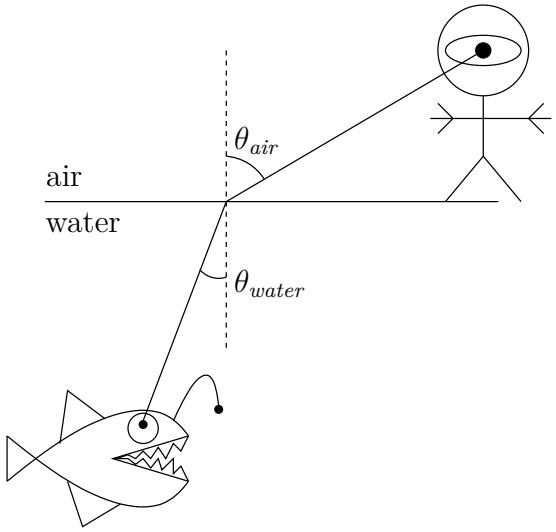


Figure 0.3: Symmetry of Snell's Law

pond looking at a fish under the water, the light rays coming from the fish to you follow the same path as the light rays going from you to the fish: light rays can travel either direction along the path shown in fig. (0.3). So if you can see the fish, the fish can see you.

Substances with higher indices of refraction are sometimes termed "optically denser." When a ray is passing from an optically denser to an optically lighter substance (that is, from higher to lower n), the ray will be refracted to a larger angle from the normal: if the $n \sin\theta$'s are to be equal, the side with the smaller n must have the larger $\sin\theta$ and hence the larger θ to compensate. In other words, in the optically lighter substance the ray will emerge along a line closer to the interface, like the ray on the air side in fig. (0.3). The larger the angle of incidence, the larger the angle of refraction and the closer the refracted ray to the interface. For a large enough angle of incidence, the angle of refraction will be 90°, so that the refracted ray ends up just skimming along the interface. This angle of incidence is called the critical angle: for larger angles of incidence, there is no angle of refraction that will satisfy Snell's law.[10] Physically, this means that none of the incident ray makes it into the second substance; all of it is reflected back into the first substance.[11] This is called total internal reflection. To determine the critical angle of in-

[10] Actually, there is an angle that will satisfy Snell's law in this case, but it is complex (that is, has an imaginary as well as a real part), so it doesn't make sense physically.

[11] Remember that when a light ray strikes an interface between two substances, in general it is partly transmitted through into the second substance *and* partly reflected back into the first substance. This is true for angles of incidence up to the critical angle: part of the ray makes it into the second substance; part is reflected back into the first. But past the critical angle, *all* of the incident ray is reflected back and *none* of it makes it through.

cidence at which total internal reflection first occurs, we need simply set the angle of refraction equal to 90° in Snell's law.

The laws of reflection and refraction are important because together they enable us to determine what happens to light as it encounters various objects. The focusing properties of the lenses in cameras, microscopes, and refracting telescopes, for example, are all determined by Snell's law. Likewise the focusing properties of reflecting telescopes are determined by the law of reflection.

While we're at it, we should note that the speed of light in a vacuum is $c \approx 3.00 \times 10^8$ m/s. In a substance of index of refraction n, it turns out that the wave is effectively slowed to speed $v = c/n$ and that the wavelength (which gets scrunched up just like cars in slow traffic) is correspondingly reduced to $\lambda_{in\ substance} = \lambda_{in\ vacuum}/n$. The frequency of the wave is unaffected.

0.2.2 Ray Diagrams & Images

A ray diagram traces light rays emanating from some particular point on an object. You should ultimately be able to trace rays for the following simple cases: light coming directly from an object to your eye; reflections off of a plane (that is, flat) mirror; light refracted at the interface between two substances; and light that passes through a thin lens.

An image is formed when light rays emanating from a single point on an object again converge to a point. Fig. (0.4) illustrates the simplest case: when you look directly at an object: the rays diverging from a point on the object are refracted (focused) by the lens of your eye so that they converge to a point on the retina in the back of your eyeball. As we will see shortly when we deal with lenses in more detail, the image formed on your retina is actually inverted (upside down) and backward (with left and right reversed), but your brain is able to interpret this image correctly by, in effect, extrapolating the rays that have struck the retina backward to their point of convergence on the object from which they originated, and that's where you perceive the

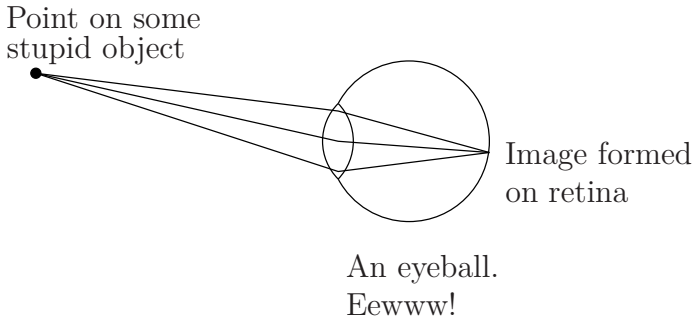


Figure 0.4: Looking Directly at an Object

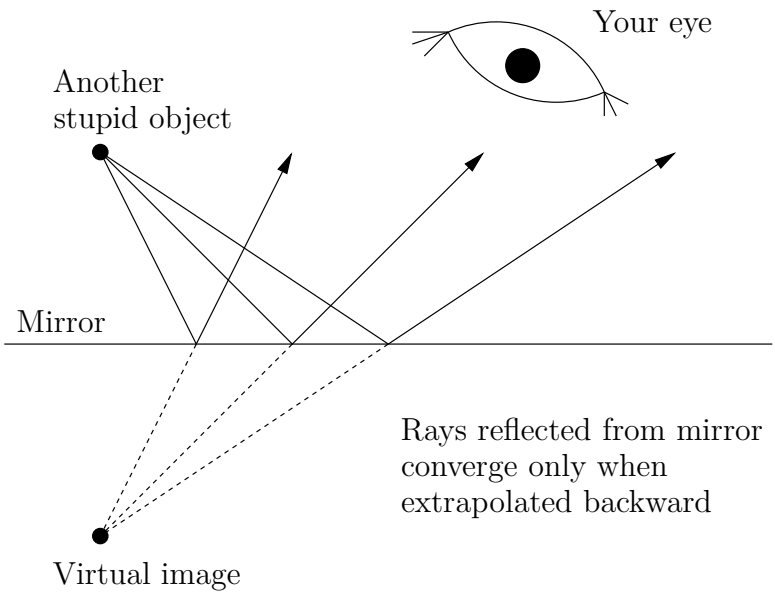


Figure 0.5: Mirror Image

object to be. Thanks to evolution, your brain is very good at doing this — it would, for example, be very unfortunate in a Darwinian sense if a snarling, hungry lion 3 m away on your right appeared to be 30 m away or to be on your left.

If someone were somehow magically able to see your retina while you were looking at the object, that person would see a tiny image of the object formed on your retina, just like the images formed on a movie screen.[12] In cases like this, when the light rays that give rise to an image actually converge to or pass through the image, so that a movie screen placed at that location would be illuminated by the image, the image is called real. If, on the other hand, the light rays converge only when extrapolated backward to a location from which they never came, the image is termed virtual. Fig. (0.5) illustrates the simplest example of a virtual image: the one formed by reflection off of a plane (that is, flat) mirror. The horizontal line in fig. (0.5) represents the mirror viewed edge-on, from above. As is familiar to everyone from looking into the bathroom mirror, the image appears to be behind the mirror, but of course the light rays do not come from there; rather, the rays come from an object in front of the mirror, are reflected from the mirror's surface, and then have to be extended backward in order to converge to a point. Since

[12] Actually, your ophthalmologist *is* able to look at your retina, but, because your retina isn't much like a movie screen, all he or she sees is a bunch of disgusting blood vessels and rods and cones. Ophthalmologists seem to be into that sort of thing. Anyway, if you overlaid your retina with a miniature movie screen, what we're saying above would hold true.

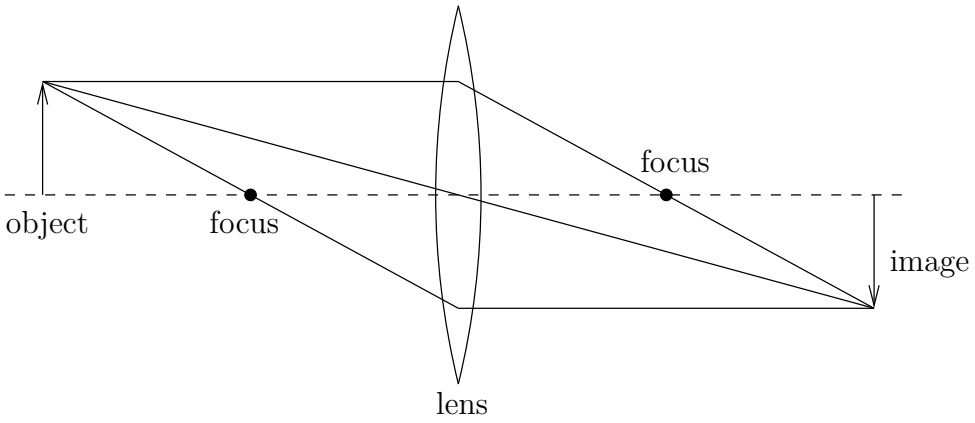


Figure 0.6: Thin Lens Rays

your brain is in the business of doing such backward extrapolations of the rays intercepted by your eye, this point of intersection is where you perceive the object to be, even though the path of the light rays from the object never passes through that location.

0.2.3 Thin Lenses

Lenses refocus light to form an image. More precisely, lenses refract light rays emanating from a point on an object in such a way that the rays converge to a common point, thus forming an image at that point of convergence. The lens of a camera, for example, focuses the rays coming from the various objects in its field of view onto corresponding points on the film, and a magnifying glass focuses the rays of the Sun onto an ant.

We will work only with a special (but very common) kind of circular lens known as a thin lens.[13] A thin lens has two defining characteristics: the two surfaces of the lens are spherical in shape, and the thickness of the lens at its center is small compared to the radii of curvature of these surfaces. Having forgone the derivation of the law of refraction, we will also forgo the derivation of the refractive properties of such lenses — the enlightenment the derivation would offer would not be worth the time and energy — but it turns out that to a reasonably good approximation:

- The lens has two foci, each focus one focal length away from the lens along its axis (that is, the perpendicular axis through the center of the lens), as shown in fig. (0.6). The focal length is the only parameter of a thin lens; all of the lens's optical properties — in particular the path

[13] There used to be an excellent interactive thin-lens ray diagram at `http://wigner.byu.edu/ThinLens/lens& mirror/lensDemo.html`, but unfortunately it seems to have disappeared. That crazy Internet!

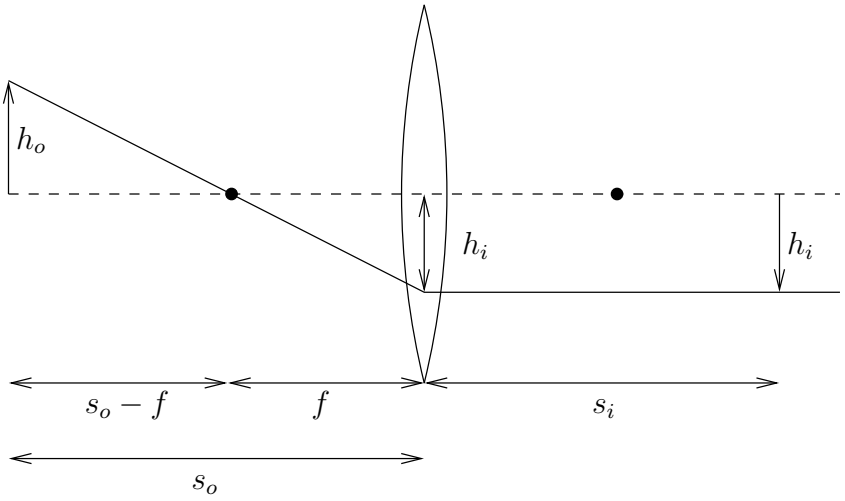


Figure 0.7: Proving the Lens Equation

of light rays through the lens — are determined by its focal length. To wit:

- Rays passing through the lens obey three geometric rules, illustrated in fig. (0.6):[14]

  1. A ray through the center of the lens is undeflected.
  2. A ray parallel to the axis of the lens is deflected so that it travels through the focus on the far side of the lens.
  3. A ray through the focus of the lens will emerge parallel to the axis of the lens on the far side.

- As we will show below, a direct consequence of the three rules for tracing rays is the lens equation:

$$\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f} \tag{0.1}$$

  where f is the focal length and where s_o and s_i are the distances (measured along the axis of the lens) from the lens to the object and image, respectively, as shown in fig. (0.7).

[14] Obviously most of the zillions of light rays that emanate from a point on an object and are intercepted by the lens do not fall into one of these three cases. But all of the rays from a common point on an object will converge to the same point after passing through the lens, so the thin-lens rules are sufficient to determine where that convergence point is located. Actually, more than sufficient; two would do.

- As we will also show below, another direct consequence of the three rules for tracing rays is a simple relation for the relative heights h_i of the image and h_o of the object shown in fig. (0.7): [15]

$$\frac{h_i}{h_o} = \frac{s_i}{s_o} \tag{0.2}$$

"Proof" of the Lens Equation

In fig. (0.7), we have drawn the ray corresponding to the last of the three thin-lens ray-tracing rules given above and noted some of the distances involved. Because the two triangles on the object (left) side of the lens are similar triangles, the ratios of their vertical to their horizontal sides will be equal:

$$\frac{h_o}{s_o - f} = \frac{h_i}{f}$$

which can be rearranged into

$$\frac{h_i}{h_o} = \frac{f}{s_o - f} \tag{0.3}$$

Now, as you can see from fig. (0.6) the rules for ray tracing, and therefore any relation derived from these rules, are symmetric under the interchange of the image and object. Interchanging image and object in eq. (0.3) gives

$$\frac{h_o}{h_i} = \frac{f}{s_i - f}$$

which, if we take the reciprocal of both sides, becomes

$$\frac{h_i}{h_o} = \frac{s_i - f}{f} \tag{0.4}$$

Equating the two expressions for h_i/h_o given by eqq. (0.3) and (0.4), we have

$$\frac{f}{s_o - f} = \frac{s_i - f}{f}$$

Multiplying both sides by $f(s_o - f)$ yields

$$f^2 = (s_i - f)(s_o - f) = s_i s_o - s_i f - f s_o + f^2$$

[15] This ratio of heights gives us some indication of the magnification, but only a crude indication, since it does not take into account how far away the object and image are. A more accurate measure of how big an object or image appears to be would be the ratio of its height to its distance from your eye, but we aren't going to get that involved in it.

which, when simplified and rearranged, reduces to

$$s_i f + f s_o = s_i s_o$$

Finally, if we divide this through by $f s_i s_o$, we obtain the lens equation (0.1):

$$\frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{f}$$

Proof of Eq. (0.2)

Returning to eq. (0.4), we have

$$\frac{h_i}{h_o} = \frac{s_i - f}{f} = \frac{s_i}{f} - 1$$

If in this we use the lens equation (0.1) to substitute

$$\frac{1}{s_o} + \frac{1}{s_i} \qquad \text{for} \qquad \frac{1}{f}$$

we obtain

$$\frac{h_i}{h_o} = s_i\left(\frac{1}{s_o} + \frac{1}{s_i}\right) - 1 = \frac{s_i}{s_o} + 1 - 1 = \frac{s_i}{s_o}$$

which is indeed eq. (0.2).

We Return to Our Regularly Scheduled Programming

Note that the ray tracing rules and the lens equation (0.1) are symmetric under the interchange of the object and image.

In the lens equation (0.1), a positive solution for the image distance s_i indicates that the image is on the opposite side of the lens from the object, which is the case illustrated by fig. (0.6). As you can see, when the object is outside the focal length of the lens you get an inverted (upside-down) real image on the far side of the lens.

A negative solution for s_i means that the image is on the same side of the lens as the object, as shown in fig. (0.8). The same three rules are used for ray tracing, but note that there is a little weirdness: since the ray going through the focus on the object's side of the lens would be going away from the lens, you have apply this rule in the slightly perverse way shown in the figure. As you can see, when the object is inside the focal length, you have to extrapolate the rays backward to get them to converge at a point, and the result is that you get an erect (right-side-up) virtual image on the same side

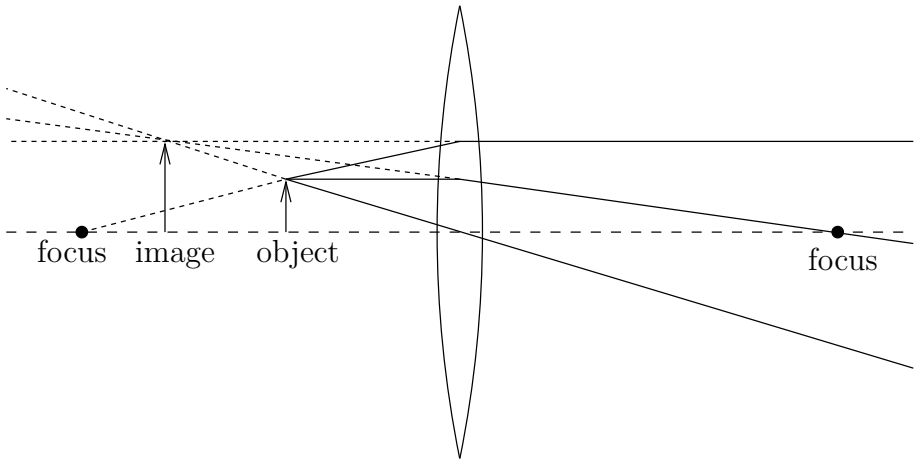


Figure 0.8: Case of an Object Inside the Focal Length

of the lens as the object — virtual because the light rays never pass through the location of the image: although when looking through the lens you would perceive (that is, literally see) an image at the location shown in fig. (0.8), if you were to place a movie screen at the image's location you would not see it on the screen.

But wait: there are still more complications. *Bwahahaha!* The thin lenses with which we have been dealing up to this point are what are known as converging lenses. There is, however, another kind of thin lens, known as a diverging lens. Fortunately, all of the results for converging lenses apply to diverging lenses as well, *with two differences*:

- *A diverging lens has an effectively negative focal length.* Thus, if you are dealing with a diverging lens of 15 cm focal length, you would use $f = -15$ cm in the lens equation.

  And because it has an effectively negative focal length,

- *For a diverging lens the roles of the two foci are reversed when you are applying the three rules for ray tracing.*

One consequence of these differences is that there is only one case for a diverging lens: whether the object is inside or outside the focal length, you always get an erect virtual image on the same side of the lens as the object, as shown in fig. (0.9). This is, in fact, the reason for the terms *converging* and *diverging*: for a converging lens (at least when the object is outside its focal length), the rays converge to a point on the far side of the lens; for a diverging lens, the rays diverge from each other on the far side.

It turns out that doubly convex lenses are always converging, doubly concave always diverging, and concavoconvex lenses can be either converging

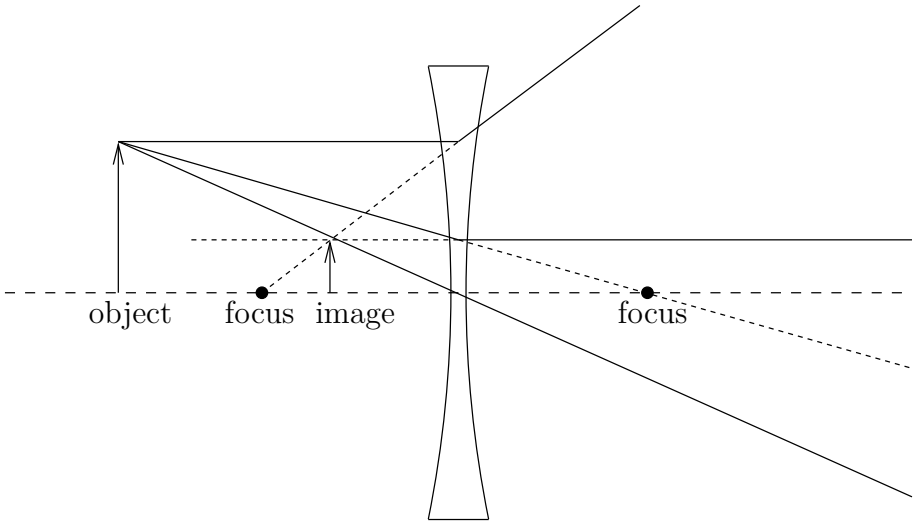


Figure 0.9: Ray Tracing for a Diverging Lens

or diverging, depending on which side has greater curvature (that is, a smaller radius of curvature for its spherical surface).[16]

0.2.4 Optical Instruments

A magnifying glass is just a converging lens: starting with the lens close to the object, you gradually move the lens as far from the object as you can without getting too much distortion. The object is then close to, but still within, the focal length, and the resulting erect virtual image, which appears on the same side of the lens as the object, is both large and distant. In addition to the benefit of the magnification, the magnifying glass makes the object seem farther away and thus easier for farsighted people to focus on.

In a two-lens microscope or telescope, the lens closest to the object is called the objective lens, the lens through which you look the eyepiece. With the microscope, the object is placed somewhat outside the focal length of the objective, so that a somewhat magnified inverted real image is formed on the opposite side of the objective. With the telescope, the object is very far away, so that a very small inverted real image is formed just outside the focal length on the opposite side of the objective. In both instruments, the eyepiece is then used as a magnifying glass, the object of which is the real image formed by the objective lens. With the microscope, both the objective and the eyepiece magnify the image; with the telescope, the objective actually

[16]There is of course a quantitative relation for the focal length of the lens in terms of the radii of the lens's surfaces and the indices of refraction of the lens and the fluid with which the lens is surrounded, but that relation is not terribly illuminating, and we will therefore limit ourselves to the qualitative observation above.

shrinks the image, which must then be magnified still more by the eyepiece in order to get an overall magnification greater than one. For this reason, the objective lens of the telescope must have a longer focal length than the eyepiece.

The above paragraph will, of course, have fallen pathetically far short of imparting to you an understanding of the microscope and telescope, but this will all become much more clear when you actually construct a microscope and telescope for yourself in the lab.

0.2.5 The Eye & Corrective Lenses

When you focus on an object, the lens of your eye is focusing the light rays coming from the object so that an image — an inverted real image — is formed on the retina at the back of your eyeball. The image distance is therefore the distance from the lens to the retina, that is, the diameter of your eyeball. Since this distance is fixed — you can't change the size of your eyeball — , the focal length of the lens of your eye must adjust in order for you to focus clearly on objects at varying distances. This is accomplished by a series of muscles around the perimeter of the lens, which contract or relax to make the lens relatively bulgy or flat, corresponding to relatively short and long focal lengths, respectively.

When you apply the lens equation to human vision, the value of the image distance will be fixed (at roughly 2.5 cm for most people):

$$\frac{1}{f} = \frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{s_o} + \frac{1}{2.5\,\text{cm}}$$

You then typically solve for the focal length f needed to focus on objects at various distances s_o. Shorter focal lengths (bulgier lenses) are needed to see closer objects, longer focal lengths (flatter lenses) to see more distant objects.

Ideally, your far point (the farthest you can see clearly) is infinity. Myopia (nearsightedness) occurs when the lens, even when fully relaxed, still has too much curvature to focus on distant objects. One common cause of myopia is spending a great deal of time doing close work like reading or working on the computer: although what your mother told you about getting stuck cross-eyed isn't true,[17] with repeated, prolonged strain the lens of the eye will actually "grow" into the bulgy shape needed for close vision.[18] To correct myopia, the lens of the eye must be combined with a diverging lens that

[17] Not to mention what she told you about swallowing seeds or bubblegum.

[18] Lest any of you conclude that the ideal solution is to do less reading, we should note that the recommendation of eye-doctors is instead that when doing a stint of reading or computer work, you take short but fairly frequent breaks to relax the eye strain by glancing out the window or at something fairly distant.

effectively "cuts out" some of the excess bulge in the eye's lens.[19]

With normal vision, your near point (the closest you can see clearly) should be about 15 cm or so. Hyperopia (farsightedness) is also known as presbyopia because of it occurs predominantly in older people. It used to be though that the cause was a weakening of the muscles around the lens of the eye with age, that the muscles became too weak to make the lens bulge enough to focus on close objects. It is now known that the muscles remain quite strong throughout life, and that the lens of the eye in fact grows bulgier with age; the real cause of hyperopia turns out to be a degradation of the refractive index (and thus focusing power) of the proteins that make up the bulk of the lens. To correct hyperopia, the lens of the eye must be combined with a converging lens that effectively increases the focusing power of the eye's lens.

It turns out that when two thin lenses are combined by putting one right after the other (as when you put on glasses or contact lenses), the combined focal length is given by

$$\frac{1}{f_{combined}} = \frac{1}{f_1} + \frac{1}{f_2}.$$

In the context of vision and corrective lenses, this becomes

$$\frac{1}{f_{with\ glasses}} = \frac{1}{f_{eye}} + \frac{1}{f_{glasses}} \tag{0.5}$$

which can be used to solve for the focal length of the glasses needed.

For example, suppose you are nearsighted and can see clearly only out to 1.5 m — that beyond 1.5 m, objects become progressively fuzzier. For an object at 1.5 m = 150 cm, the lens equation, applied to your eye, gives

$$\frac{1}{f} = \frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{150} + \frac{1}{2.5}$$

or $f = 2.46$ cm for the focal length of your unaided eye. To see clearly out to infinity would require

$$\frac{1}{f} = \frac{1}{s_o} + \frac{1}{s_i} = \frac{1}{\infty} + \frac{1}{2.5}$$

or $f = 2.5$ cm for the focal length you want to attain with your glasses on. The relation (0.5) for combining lenses thus gives

$$\frac{1}{f_{with\ glasses}} = \frac{1}{f_{eye}} + \frac{1}{f_{glasses}}$$

[19] Laser surgery to correct myopia literally cuts out some of the excess bulge, so that the relaxed lens is flat enough for you to be able to see clearly out to infinity.

$$\frac{1}{2.5} = \frac{1}{2.46} + \frac{1}{f_{glasses}}$$

which yields $f_{glasses} = -150\,\text{cm}$ for the focal length of the glasses that will fully correct your nearsightedness.[20]

Optometrists conventionally specify, not the focal length of the corrective lenses, but the inverse (that is, the reciprocal) of the focal length, and they do this in units of diopters. 1 diopter $= 1\,\text{m}^{-1}$, so that having a prescription of -2 diopters means that for your glasses

$$\begin{aligned}\frac{1}{f} &= -2\,\text{diopters} \\ &= -2\frac{1}{\text{m}} \times \frac{1\,\text{m}}{100\,\text{cm}} \\ &= -\frac{1}{50\,\text{cm}} \\ f &= -50\,\text{cm}\end{aligned}$$

And, going the opposite direction, the $f_{glasses} = -150\,\text{cm} = -1.5\,\text{m}$ that we obtained in the preceding example would be $1/(-1.5) = -0.67$ diopters. The correction for myopia or hyperopia is called the spherical correction, and, in addition to being specified in diopters, is broken into separate prescriptions *OD* and *OS* (*oculus dexter* and *oculus sinister*) for the right and left eyes, respectively.

Astigmatism is like myopia and hyperopia, but is not spherically symmetric: with astigmatism, the inability to focus is along some particular axis in the plane of the lens of your eye, and the direction of this axis must also be specified in the prescription. This part of the prescription is called the cylindrical correction, and lenses that correct for astigmatism must therefore be correctly oriented over your eye. This is not a problem for glasses, but requires special weighting for contact lenses.

0.3 Wave Optics

0.3.1 Interference, Diffraction, Dispersion, & Polarization

We don't want to get involved in a quantitative treatment of light as a wave at this point, but there are a few effects with which you should be familiar qualitatively.

[20] Those of you who followed the algebra closely will have noted that it is no accident that we have ended up with $f_{glasses}$ equaling the negative of your far point. But we want you to be able to reproduce the steps in the logic, not just fire back a canned result. And for corrections to hyperopia, there are no such shortcuts.

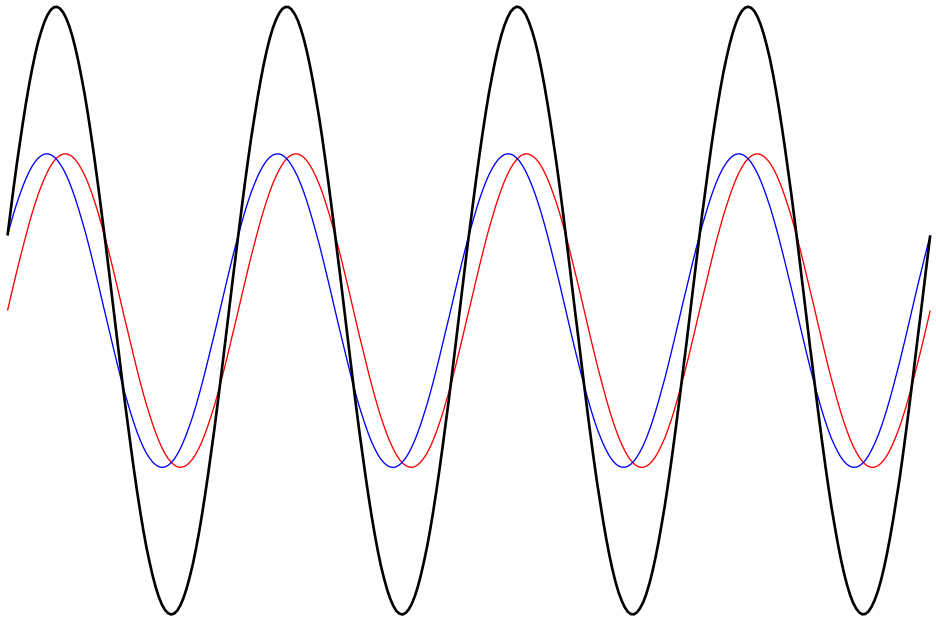


Figure 0.10: Wave Superposition: Mostly Constructive

When waves, such as ripples from tossing pebbles into a pond, meet, they combine additively. That is, the net or resultant wave is simply the arithmetic sum of the individual waves. This combining of waves is called superposition or interference[21] and applies to light just as to water waves. If, for example, one wave is at a crest while the other is at a trough, they will at least partially cancel each other out, while if they are aligned crest-to-crest and trough-to-trough, they will reinforce each other to produce a bigger wave. Fig. (0.10) shows the superposition of two waves (red and blue) that are identical except that one is slightly shifted (out of phase) relative to the other; in this case the waves largely reinforce each other and produce a resultant (black) nearly twice their size. Fig. (0.11) shows the superposition of two similar waves, but now the shift (phase difference) is large enough that when one is at a crest, the other is near a trough and vice versa; in this case the waves largely cancel each other and produce only a small resultant. And fig. (0.12) shows the superposition of two waves that have the same amplitude but not quite the same wavelength, so that their phase relationship varies: at

[21] An excellent interactive animation of wave superposition can be found at `http://www.phy.ntnu.edu.tw/java/waveSuperposition/waveSuperposition.html`. (If you alter the numbers in an edit-field at the top of this applet, you have to hit ENTER to make the change effective.) You might also check out `http://www2.biglobe.ne.jp/~norimari/science/JavaEd/e-wave2.html` and `http://www2.biglobe.ne.jp/~norimari/science/JavaEd/e-wave3.html`.

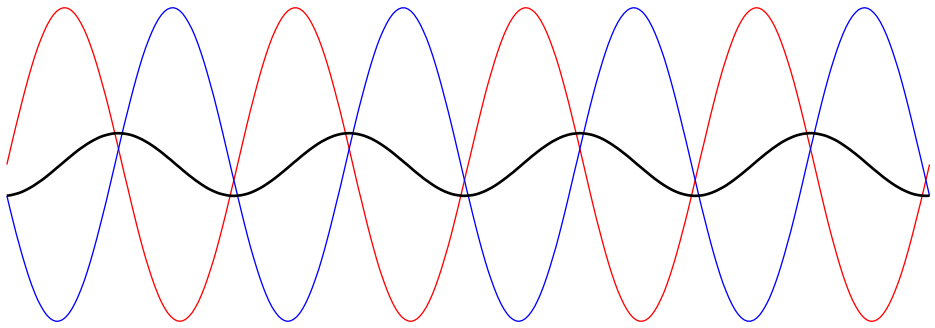


Figure 0.11: Wave Superposition: Mostly Destructive

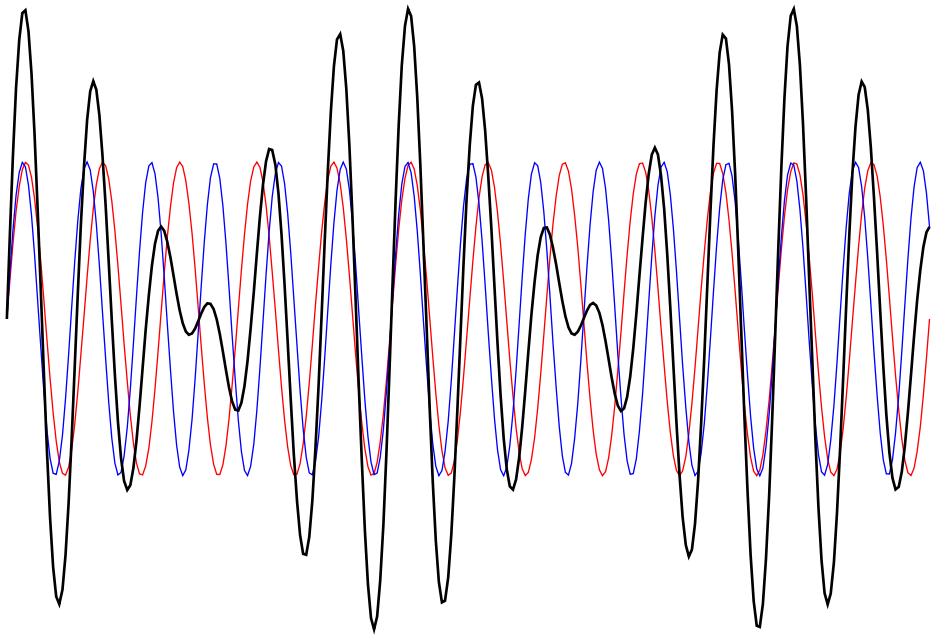


Figure 0.12: Wave Superposition: Beats

some points the waves will be nearly aligned (crest to crest), at others nearly antialigned (crest to trough), so that they oscillate between reinforcing and canceling each other. The result is a series of "beats" in the resultant wave.

It is interference that causes the rainbows on oil slicks, soap bubbles, and other thin films: of the light incident on the film, part is reflected from the near surface of the film, part is transmitted into the film. Of the transmitted light, part will be reflected upon reaching the far side of the film and travel back to the near side of the film — where the process of partial reflection and partial transmission will be repeated, *ad nauseam*. The result is that, of the light reflected back from the film, part comes from the initial reflection when the light first hit the film's near surface and part comes from the myriad bounces that occur inside the film. In this recombining of the light with itself, some wavelengths (the colors you see) interfere with themselves constructively, others (the colors you don't see) destructively. Which wavelengths interfere constructively and which destructively depends on the thickness and index of refraction of the film. When you see rainbow effects, it is because the thickness of the film varies, and with it the wavelength and color that is most strongly reflected.

To determine what happens when a light wave encounters an object — some sort of blob, an opaque wall with a small hole or slit cut in it, or whatever — , we would need to apply the Maxwell equations, fitting the wave solution of these equations to boundary conditions corresponding to the object.[22] We would then find that as a result of encountering the object the wave was altered, often into a pattern of alternating high and low intensity known as a diffraction pattern. Diffraction effects are of significant magnitude only when the objects causing the diffraction are comparable in size to the wavelength of the light. It is diffraction around buildings, etc., that causes the alternating regions of better and worse radio and TV reception in your dorm rooms.[23]

Although the value of the index of refraction of course differs from substance to substance — it has different values for air, water, glass, etc. — for any one substance we have treated it as a constant. In fact, the index of refraction depends on the wavelength of the light being refracted, so that the angle of refraction varies with the wavelength. This variation is generally very slight, but it is enough to give rise, among other effects, to the "rainbow"

[22] Introductory textbooks often invoke Huygen's principle to explain what happens. This is unfortunate, because this approach, while strictly speaking not incorrect, is misleading and hand-wavy and obscures what is really happening at a fundamental level. You might as well invoke hordes of invisible gremlins.

[23] To be clearly observable, diffraction effects require a coherent, monochromatic wave, that is, a single wave of a single color (such as a laser produces). Ordinary daylight is a royally jublified time-varying superposition of different waves of different colors, so that although diffraction still occurs, the resulting "pattern" is an equally jublified mess.

separation of the colors in visible light by a glass prism that is, thanks to Pink Floyd, familiar to you all. This separation of wavelengths by refraction is called dispersion. Dispersion is also responsible for the separation of colors seen in rainbows: rainbows are the result of sunlight being refracted into, totally internally reflected inside of, and then refracted back out of, droplets of water in the atmosphere. The dispersion occurs at the refractions on the way into and back out of the drop.

Light is a wave oscillation of electric and magnetic fields. More precisely, the electric-field oscillation, magnetic-field oscillation, and direction of propagation are all mutually perpendicular, like a set of xyz axes. The direction along which the electric-field oscillation takes place is called the axis of polarization and determines how the light wave interacts electrically with substances. Usually, the polarization of a light wave makes little difference. Reflected light, however, tends to be polarized more or less strongly parallel to the reflecting surface. Polarizing filters allow only one direction of polarization to pass through, and polaroid sunglasses are therefore particularly effective against glare.

0.3.2 Why the Sky is Blue

Why is the sky blue?

Why not?

Okay, so you want a more satisfying answer. But before you read on, consider Bokonon's advice:[24]

> Beware of the man who works hard to learn something, learns it, and finds himself no wiser than before. He is full of murderous resentment of people who are ignorant without having come by their ignorance the hard way.

So our advice is to just put the book down now. But if you simply must, go ahead, read on. Just don't say we didn't warn you.

A proper explanation would require quantum calculations that we are obviously in no position to pursue, and although it turns out that a crude sort of semiclassical approximation will be good enough to yield the effect, this will still require us to make use of some concepts and results many chapters ahead of where we are now. But we can't very well pass over so basic a question as why the sky is blue, and if we are going to address that question, this is the place to do it.

[24] From Chapter 24 of Vonnegut's *Cat's Cradle*. You might think this is a little melodramatic in the current context, but then you haven't seen what's coming.

Since sunlight is strong, we will treat it as a continuous electromagnetic wave (as opposed to the collection of discrete quantum particles, known as photons, that characterize dim light, as discussed in Chapter 22). As this electromagnetic wave passes through the atmosphere, it exerts electromagnetic forces on the electric charges (the electrons and nuclei) in the molecules of the atmospheric gases. The electrons, being much lighter than the nuclei, respond more strongly to these forces than do the nuclei. If you read on ahead into the chapters on kinematics, dynamics, and electromagnetism, then invent a time machine and return to the current moment,[25] you will recall that the electric force F exerted by an electric field E on a charge q is given by

$$F = qE \tag{0.6}$$

For the electromagnetic oscillation that constitutes an electromagnetic wave, the electric field is of the sinusoidal form

$$E = E_0 \cos \Omega t \tag{0.7}$$

where t is the time, E_0 is the field's amplitude (that is, its peak value), and Ω is its angular frequency. (This angular frequency Ω is related to frequency f by $\Omega = 2\pi f$: Ω and f really measure the same thing — how rapidly the wave is oscillating — , it's just that Ω measures that rate in terms of angle rather than cycles, and there are 2π radians in each cycle.) Using eq. (0.7) in (0.6), we have for the electric force felt by the electrons

$$F = qE = qE_0 \cos \Omega t = F_0 \cos \Omega t \tag{0.8}$$

where we have defined $F_0 = qE_0$. Now, the electrons are in stable molecular orbits of a quantum in nature, but we can crudely approximate that, as is shown in Chapter 9 quite generally for objects in classical stable equilibria, they will respond in a spring-like way to this electric force. And the application to a spring of a sinusoidal "driving" force like that of eq. (0.8) just happens to be exactly the case worked out in §9.6. Well, almost exactly; in the present context we have no need of the phase φ or the damping parameter β that were included in §9.6. But if we drop that excess baggage and use the definition $a = d^2x/dt^2$ for acceleration that you learned during your time travels, the result (9.25) of §9.6 yields

$$a = \frac{d^2}{dt^2}\left(\frac{F_0/m}{\sqrt{(\omega_0^2 - \Omega^2)^2}} \cos \Omega t \right)$$

[25] We did warn you to put the book down, didn't we? Anyway, if you have no taste for the technical details, you can skip ahead to just after eq. (0.11)without losing your sanity or feeling too guilty. The calculations will make much more sense if you come back to them later in the course.

$$= -\frac{F_0}{m}\frac{\Omega^2}{|\omega_0^2 - \Omega^2|}\cos\Omega t$$
$$= -\frac{qE_0}{m}\frac{\Omega^2}{|\omega_0^2 - \Omega^2|}\cos\Omega t \tag{0.9}$$

where we have used $F_0 = qE_0$ and where ω_0 is the natural angular frequency associated with the orbit of the electrons. It turns out that the angular frequency Ω of visible light is small enough compared to ω_0 that we can set $\omega_0^2 - \Omega^2 \approx \omega_0^2$ in the denominator of eq. (0.9). Thus

$$a \approx -\frac{qE_0}{m}\frac{\Omega^2}{\omega_0^2}\cos\Omega t \tag{0.10}$$

Having arrived at a result for the acceleration that the electrons experience as a result of the passing light waves, we now pull our biggest rabbit yet out of the hat: it turns out that the power P radiated by a charge q undergoing an acceleration a is given by

$$P = \frac{2}{3}\frac{q^2a^2}{c^3}$$

where c is the speed of light. With the acceleration of eq. (0.10), this means that the power radiated by the electrons is

$$P \approx \frac{2}{3}\frac{q^2}{c^3}\left(-\frac{qE_0}{m}\frac{\Omega^2}{\omega_0^2}\cos\Omega t\right)^2 = \frac{2q^4E_0^2\Omega^4}{3m^2c^3\omega_0^4}\cos^2\Omega t \tag{0.11}$$

Quite a mess, you might think. But the key thing to note is the dependence on the angular frequency Ω of the light wave: $P \sim \Omega^4$, so that the power radiated increases dramatically with the frequency of the light wave. Thus visible light toward the high-frequency (blue) end of the visible spectrum is more strongly scattered by the gas molecules of the atmosphere than light toward the low-frequency (red) end. And that has consequences:

If there were no atmosphere, light from the Sun would travel straight in to the Earth's surface. When you looked around you, you would see your surroundings because of the sunlight reflected by them. And when you looked at the Sun, you would see the familiar bright yellowy ball [26] because your eyes would be catching the light coming straight from the Sun to you. But when you looked out anywhere else in the sky, you would see only blackness and stars — the only light coming at you from those parts of the sky would be starlight. But since we do have an atmosphere, there is light coming at you from these other parts of the sky: the light scattered by the gas molecules of the atmosphere. The colors that dominate this light are those that are most strongly scattered, which are at the blue end of the visible spectrum.

[26] At least, you would for a very short while, until your retina burned out and you went blind.

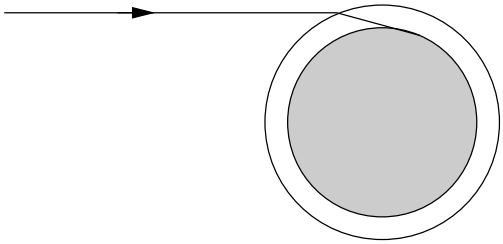


Figure 0.13: The Path of Sunlight at Sunrise & Sunset

This same effect is responsible for the red color of sunrises and sunsets: When the Sun is more or less directly overhead, the atmosphere, while it scatters a great deal of the bluer light, isn't deep enough to scatter away all the blue light, with the result that the Sun still looks more or less white.[27] But at sunrise and sunset, the light you are seeing when you look at the Sun is following a much longer path through the atmosphere, both because of the glancing angle and because of refraction, as very crudely shown in fig. (0.13). Of course, the path in fig. (0.13) doesn't look all that much longer, but the proportional height of the atmosphere has been greatly exaggerated in the figure; with a shallower, more realistically scaled atmosphere, the path would be a lot longer — long enough that most of the bluer light is scattered away and the light that does make it through is predominantly in the redder end of the visible spectrum.

0.4 Parabolic Mirrors

As we will show below, a parabolic mirror (meaning a mirror in the shape of the surface of revolution you get when you rotate a two-dimensional plot of a parabola) focuses rays that come in parallel to its axis at a point. This property makes parabolic mirrors useful for building telescopes: because they come from a source so far away, the incoming light rays from stars are very nearly parallel to each other, and if the parabolic mirror's axis is aligned with the rays, those rays will be focused at a point just as they would by a lens. In fact, reflecting telescopes made with parabolic mirrors have an important advantage over refracting telescopes made with solid lenses: when the starlight is very dim, the cross-sectional area of the telescope must be large in order to collect enough light for the star to be seen, and it is much cheaper to build large mirrors than large lenses.

[27] White being the color created by superposing all the various colors of the spectrum, as you all know well in these days of RGB computer monitors. Next time you are at your computer, take a very close look at the screen, preferably with a magnifying glass, so that you can see the individual pixels: to create white, the monitor simply illuminates clusters of adjacent red, green, and blue pixels.

Parabolic mirrors are also used in spotlights: since the paths of light rays are always reversible, rays emanating from a light source at the focal point of a parabolic mirror will end up coming out of the mirror parallel to its axis. Thus a strong light source at the focus of the parabolic mirror will result in a bright, well-collimated beam of nearly parallel outgoing rays. You may have seen flashlights that have a pliable metal reflector, the curvature of which can be adjusted by turning a collar around the flashlight's front lens: as the curvature varies from more to less parabolic, the beam coming out of the flashlight flares from narrow to wide.

First we will prove that rays coming in parallel to the axis of a parabolic mirror will be reflected through a common point on the axis of the mirror, without worrying about whether mirrors of other shapes have this property as well. It will be sufficient to show that two-dimensional parabolic mirrors have this property; if so, then clearly three-dimensional parabolic mirrors generated by rotating the two-dimensional mirrors about their axes will as well. The general expression for a two-dimensional parabola is $y = ax^2 + bx + c$,[28] but without loss of generality we can, for simplicity, align the axis of our parabola with the y axis and put its vertex at the origin, as shown in fig. (0.14). If we do so, then $b = c = 0$ and the expression for the parabola reduces to $y = ax^2$. Now, suppose a light ray comes in parallel to the y axis, as shown by the blue line in fig. (0.14). This ray will strike the mirror at some general point $(x, y) = (x, ax^2)$. To determine where the reflected ray will pass through the axis of the mirror we need to work through the geometry shown in fig. (0.14): we know that the point (x, ax^2) lies at one end of the reflected ray, and if we can determine the slope of the reflected ray from the various angles shown in fig. (0.14), then we can figure out the point $(0, y_0)$ where the reflected ray will pass through the y axis. To accomplish this, recall (see fig. (0.15)) that the slope of the tangent to a curve, dy/dx, is the tangent of the angle θ that the tangent line makes with the horizontal axis:

$$\tan\theta = \frac{dy}{dx}$$

In fig. (0.14), the green line is the normal to the mirror. That is, the green line is perpendicular to the tangent to the mirror (the red line). From fig. (0.14) you can see that, if θ is the angle between the tangent and the horizontal, it is also both the angle between the incident ray and the normal and, by the law of reflection, the angle between the reflected ray and the normal. From

[28] Actually, this expression assumes that the axis of the parabola is parallel to the y axis; a more general expression would be something like

$$\cos\varphi\, y - \sin\varphi\, x = a(\sin\varphi\, y + \cos\varphi\, x)^2 + b(\sin\varphi\, y + \cos\varphi\, x) + c$$

But let's not be ridiculous.

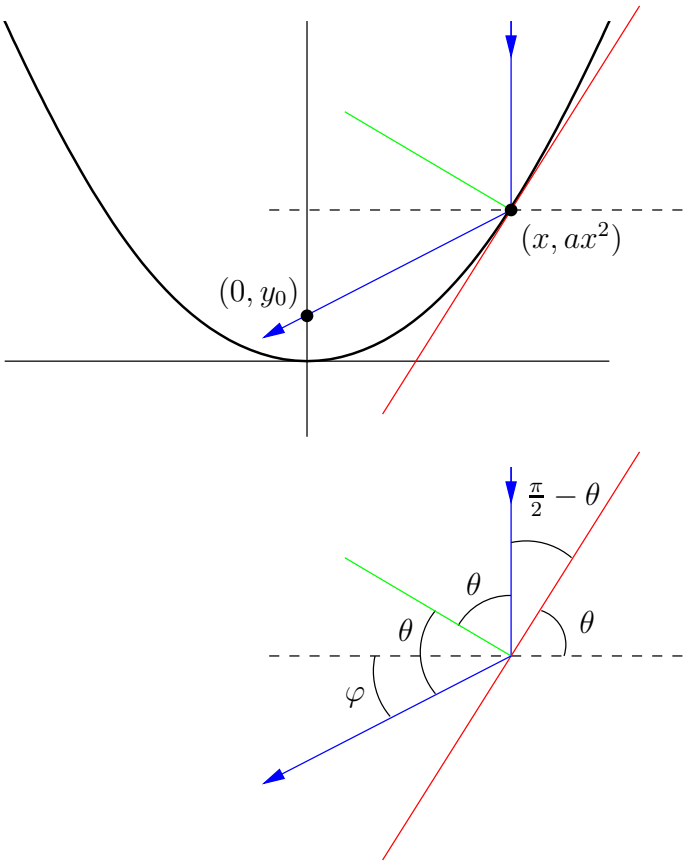


Figure 0.14: A Ray Striking a Parabolic Mirror

fig. (0.14) we thus have

$$\varphi + \pi = \theta + \theta + \left(\frac{\pi}{2} - \theta\right) + \theta$$

which simplifies to

$$\varphi = 2\theta - \frac{\pi}{2} = -\left(\frac{\pi}{2} - 2\theta\right)$$

So the slope of the reflected ray is given by

$$\tan\varphi = -\tan\left(\frac{\pi}{2} - 2\theta\right)$$

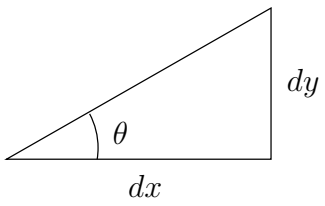


Figure 0.15: The Geometry of Tangents

Our immediate task is to relate this to $\tan\theta$, the value of which we do know:

$$\tan\theta = \frac{dy}{dx} = \frac{d(ax^2)}{dx} = 2ax$$

With a little trigonometric gymnastics, we obtain

$$\begin{aligned}
\tan\varphi &= -\tan\left(\frac{\pi}{2} - 2\theta\right) \\
&= -\cot 2\theta \\
&= -\frac{\cos 2\theta}{\sin 2\theta} \\
&= -\frac{\cos^2\theta - \sin^2\theta}{2\sin\theta\cos\theta} \;\; \frac{\frac{1}{\cos^2\theta}}{\frac{1}{\cos^2\theta}} \\
&= -\frac{1 - \tan^2\theta}{2\tan\theta} \qquad (0.12) \\
&= -\frac{1 - (2ax)^2}{2(2ax)} \qquad (0.13)
\end{aligned}$$

Since the reflected ray passes through the points $(0, y_0)$ and (x, ax^2), its slope is also given by

$$\frac{ax^2 - y_0}{x - 0} \qquad (0.14)$$

Equating our two expressions (0.13) and (0.14) for the slope of the reflected ray, we have

$$\frac{ax^2 - y_0}{x} = -\frac{1 - (2ax)^2}{2(2ax)}$$

which, when solved for the value y_0 where the reflected ray crosses the y axis, yields

$$y_0 = \frac{1}{4a}$$

Since this result for y_0 is independent of x — that is, independent of where the incoming ray strikes the parabolic mirror — , this proves that all incident rays parallel to the mirror's axis are focused at a common point on the mirror's axis. For the parabola $y = ax^2$ that point happens to be a distance $1/4a$ from the base of the mirror.

We can also pursue the proof in the opposite direction, starting from the assumption that the incident rays are focused at the point $(0, y_0)$ for a mirror that passes through the origin, and then asking what shape of mirror (that is, what function $y(x)$) will make this happen. This proof is very similar to what we just did above; the difference is just that now we are no longer

assuming $y = ax^2$, so that we must express the point where the incident ray strikes the mirror as (x, y), where $y = y(x)$ is as yet unknown. So our expression (0.12) for the slope of the reflected ray is still valid,

$$\tan\varphi = -\frac{1 - \tan^2\theta}{2\tan\theta}$$

but for $\tan\theta$ we have only $dy/dx = y'$, with y an as yet unknown function of x:

$$\tan\varphi = -\frac{1 - y'^2}{2y'} \tag{0.15}$$

Similarly our other relation (0.14) becomes

$$\frac{y - y_0}{x - 0} = \frac{y - y_0}{x} \tag{0.16}$$

Equating our two expressions (0.15) and (0.16) for the slope of the reflected ray now gives us a differential equation that we can in principle solve for $y(x)$:

$$\frac{y - y_0}{x} = -\frac{1 - y'^2}{2y'}$$

or, if we clear the denominators and clean up a bit,

$$x(y'^2 - 1) = 2y'(y - y_0) \tag{0.17}$$

Unfortunately, eq. (0.17) is a nonlinear differential equation, and while there are general techniques for solving linear differential equations, there are no such general techniques for nonlinear differential equations. But in the case at hand there is still a relatively simple method by which we can arrive at a solution: if we assume that there is a valid series expansion for $y(x)$ about $x = 0$, then

$$y = \sum_{n=-\infty}^{+\infty} k_n x^n$$

where the k_n are as yet unknown coefficients for which we must solve. To make this task easier, we can first ask whether there are any constraints that will restrict the range of n to something more manageable than $-\infty < n < +\infty$. And indeed there are: first of all, we can't have $y(x)$ blowing up at $x = 0$ — that clearly isn't going to give us the behavior we want for the mirror. So all $n < 0$ are ruled out. And if we now try out a term like $k_n x^n$ in eq. (0.17) by substituting [29]

$$y = k_n x^n$$

[29] We are being a bit sloppy here: really we should substitute all of the remaining series

and the corresponding

$$y' = nk_n x^{n-1}$$

we have

$$\begin{aligned} x\Big((nk_n x^{n-1})^2 - 1\Big) &= (2nk_n x^{n-1})(k_n x^n - y_0) \\ n^2 k_n^2 x^{2n-1} - x &= 2nk_n^2 x^{2n-1} - 2nk_n y_0 x^{n-1} \\ n(n-2)k_n^2 x^{2n-1} + 2nk_n y_0 x^{n-1} - x &= 0 \end{aligned} \tag{0.18}$$

where we have expanded, combined like terms, and neatened up a bit.

At this point, we need to keep in kind that eq. (0.18) is an equation for k_n in terms of the parameter n, and that *eq. (0.18) must hold for all x*. If we try out $n = 0$, eq. (0.18) gives $x = 0$, which certainly isn't valid for all x and therefore doesn't work.

If we try out $n = 1$, eq. (0.18) gives [30]

$$-k_1^2 x + 2k_1 y_0 - x = 0 \qquad \forall\, x$$

or

$$-(1 + k_1^2)x + 2y_0 k_1 = 0 \qquad \forall\, x$$

Since the terms with x^1 and x^0 vary differently as x varies — mathematically, one says that x^1 and x^0 are functionally independent — , this equality can hold for all x only if we separately have both

$$1 + k_1^2 = 0 \qquad \text{and} \qquad 2y_0 k_1 = 0$$

Even if y_0 were zero, this would still leave us with an imaginary value for k_1. So $n = 1$ doesn't work, either.

If $n \geq 3$, the first term of eq. (0.18) will involve different power of x than the other two terms and therefore be functionally independent of them, so that we will have to have the separate relation

$$n(n-2)k_n^2 = 0$$

---

for y,

$$y = \sum_{n=0}^{+\infty} k_n x^n$$

and not just individual terms in this series. We aren't going to do this because it would make the calculations much more complicated and messy than you could be expected to deal with at this point, but the proof when the substituting the whole series would not differ in character from that we are giving for the individual terms. And at any rate we never proved that a valid series expansion existed for $y(x)$, anyway.

[30] If you haven't seem it before, the symbol $\forall$ means *for all*. We suppose we could just write *for all*, but $\forall$ looks so much cooler.

which requires $k_n = 0$. This kills all terms for which $n \geq 3$.

That leaves $n = 2$ as the only possibility. For $n = 2$, eq. (0.18) gives

$$4k_2 y_0 x - x = 0$$

or

$$k_2 = \frac{1}{4y_0}$$

This works, so that we have as the only possibility for a mirror that will focus parallel incoming rays at a point the parabola $y = k_2 x^2$, with $k_2 = \frac{1}{4y_0}$. And comparing this to our result $y_0 = 1/4a$ from the proof that started with the parabola $y = ax^2$, we see that

$$k_2 = \frac{1}{4y_0} = \frac{1}{4\frac{1}{4a}} = a$$

Word.

0.5 Problems

1. Suppose you have a plate glass window 2.0 cm thick, with index of refraction 1.5.
   (a) If a ray of light is strikes the window at 20° *to its surface*, at what angle does the ray travel through the glass?
   (b) At what angle does that ray re-emerge into the air on the far side?
   (c) Does your answer to the preceding part depend on the index of refraction of the glass?

2. Back in his bachelor days, the ceiling of Gaffney's crib, which was 3.0 m above the floor, was covered with a large mirror. His 16-head disco helicopter-smoke generator hung 0.50 m below the ceiling and 1.5 m above his lava lamp (right next to the velvet picture of Elvis). If you lay on his leather sofa, 2.5 m horizontally away from the lava lamp and level with it, and took a picture of the image of the lava lamp in the mirror, what would the lens barrel read as the distance to the image? (This problem is not hard: just draw the picture.)

3. You have no doubt noticed that two, but only two, of the three spatial directions seem reversed when you look into mirror: right and left seem reversed, as do near and far, but up and down do not. That is, when you raise your right hand, your mirror image raises its left; when you move your hand toward the mirror, your mirror image moves its hand in the opposite direction (toward you); but when you raise your hand, your mirror image also raises its hand.
   (a) Explain why up and down are not reversed.
   (b) Is the apparent reversal of near and far real or illusory? That is, does the law of reflection actually reverse near and far, or is this just a matter of human perception? If the former, explain how the reversal occurs; if the latter, explain the illusion in terms of human perception.
   (c) Is the apparent reversal of right and left real or illusory? If the former, explain how the reversal occurs; if the latter, explain the illusion in terms of human perception.
   (d) Suppose you now put two mirrors at right angles, so that you generate three images: an image in each of the two mirrors, and a third image, what you might call the image of the image, formed by rays that reflect off of *both* mirrors.
      i. Draw a ray diagram, tracing and extrapolating enough rays to locate each of these three images.
      ii. Will near and far seem reversed in the "third" image?
      iii. Will right and left seem reversed in the "third" image?

4. (a) If you are on a boat out in a lake, how will refraction and reflection affect your view of objects under the water? That is, where will the objects appear to be relative to their actual locations? Consider objects everywhere, from very near the boat to very far from it.

   (b) If you are swimming underwater, how will refraction and reflection affect your view of objects above the water? That is, where will the objects appear to be relative to their actual locations? Consider objects everywhere, from directly overhead to very far away.

   (c) Did you remember to take into account the possibility of total internal reflection?

   (d) What consequences does all this have for bow- and spear-fishing?

   (e) If you were bow- or spear-fishing, are there any circumstances under which you would be able to see the fish but the fish unable to see you — or vice versa?

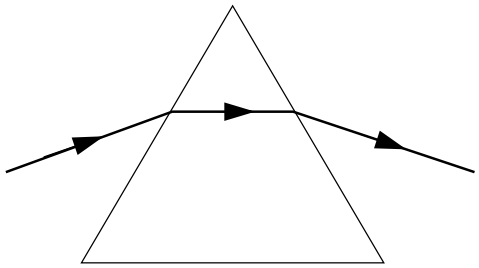


Figure 0.16: Problem 5

5. A prism of equilateral cross section has an index of refraction of 1.5. At what angle must a ray strike the prism if the path of the light inside the prism is to be parallel to its base, as sketched in fig. (0.16)?

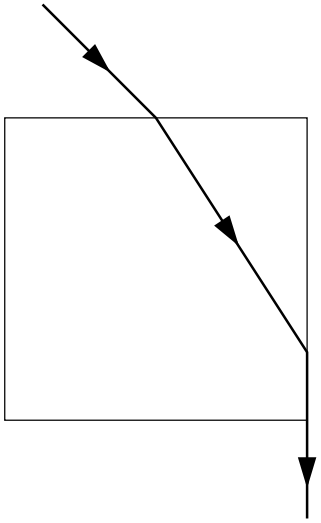


Figure 0.17: Problem 6

6. Suppose you have a translucent rectangular block of unknown index of refraction.

   (a) Playing with the block and a laser, you discover that when a light ray traveling through the air strikes the face of the block at a 45° angle of incidence, that ray is totally internally reflected (critical case) when it reaches the adjacent face of the block. That is, the path of the ray is as sketched in fig. (0.17). What is the index of refraction of the block?

   (b) At what angle will a ray incident on a face of the block at 80° re-emerge into the air from the adjacent face?

7. Suppose you have a converging lens of 15 cm focal length. When the object is 5.0, 10, 20, and 30 cm from the lens, determine

   (a) The image location.

   (b) Whether the image is real or virtual.

   (c) Whether the image is erect or inverted.

   (d) The relative heights of the image and object.

8. Do the preceding problem for the case of a diverging lens of the same focal length.

9. On a nice, sunny day, you amuse yourself by burning up ants with a magnifying glass of 15 cm focal length. How far from the ants should you hold the magnifying glass for best effect? (Don't even pretend that you didn't do this when you were a kid.)

10. (a) Why can't you get the image produced by a magnifying glass to appear on a screen, the way you can the image produced by a movie projector?

    (b) If you can't project the image produced by a magnifying glass onto a screen, how can you even see it?

11. For the yearbook, you photograph one of the dining hall's typical 15 cm roaches. If you use a 50 mm focal-length lens and the distance *between the roach and the film* is 0.50 m,

    (a) How far from the film must the lens be in order that the image be in focus on the film?

    (b) How big will the image of the roach be on the film?

    (c) Is the image of the roach on the film real or virtual, erect or inverted?

12. (a) A person with healthy eyesight can focus on objects ranging from infinity to about 15 cm away. If the distance from the lens of the eye to the retina is 2.5 cm, what range of focal lengths can the healthy eye span?

    (b) In contrast, Gaffney's eyesight is slightly worse than that of a gopher. If Gaffney's nearsightedness does not allow him to focus on objects farther away than 50 cm, what is the corresponding limit on the focal length of the lens of his eye?

    (c) What eyeglass prescription, in diopters, will fully correct Gaffney's vision? Will the glasses have converging or diverging lenses?

    (d) As Gaffney grows old, his eye muscles become so weak that he is no longer able to focus on objects *closer* than 50 cm. What eyeglass prescription, in diopters, will fully correct Gaffney's farsightedness so that he will be able to read print held 25 cm away? Will the glasses have converging or diverging lenses?

    (e) In fact, Gaffney's prescription is -2.50 diopters for the right eye and -2.00 diopters for the left eye. If Gaffney takes off his glasses and covers his left eye, at what distance will objects start to become fuzzy?

    (f) Still worse, Gaffney wears bifocals. If the focal length of the bifocal part of his glasses is -0.50 diopters, how far away can he see clearly through his bifocals? (For simplicity, use the numbers of #12b.)

    (g) At one time, Gaffney wore contacts to correct his distance vision.

        i. If Gaffney had done a lot of reading while wearing these contacts, explain how this might have adversely affected his vision.

        ii. To prevent this adverse effect, Gaffney wore a pair of reading glasses (*converging* lenses) over his contacts. Explain how this was beneficial.

13. (a) If you wear glasses or contacts and know your prescription, determine from the prescription how close up or far away, as the case may be, you can see clearly without your glasses on. Does this gibe with observation?

    (b) If you don't have four eyes but also don't have perfect vision, ascertain by observation how close up and far away you can see clearly, and determine what prescription, in diopters, would enable you to see clearly from 15 cm to infinity.

    (c) If you have perfect vision, try to find someone who does not and do #13a or #13b for his or her vision.

14. Purely out of concern for your neighbors' wellbeing, you construct a homemade telescope to spy on their swimming pool, which is 30 m away. The objective and eyepiece lenses have focal lengths of 60 and 15 cm, respectively, and the eyepiece is 76 cm from the objective lens when the telescope is focused on the neighbors.

    (a) Where, measured from the objective lens, is the image produced by the objective?

    (b) How many times smaller is the size of the image produced by the objective lens than the original object?

    (c) Where, measured from the eyepiece, is the image produced by the eyepiece?

    (d) How many times larger is the size of the overall image (that is, the image produced by the eyepiece) than the original object?

15. Though most real microscopes are of course much more sophisticated, the orientation of the image they produce is the same as that of the simple two-lens microscope discussed in §0.2.4.

    If you have made a slide with some pond water and the volvox you are watching skitters off the left side of your view, which way should you move the slide to get it back in view?

16. You examine the fuzz balls between your toes through a homemade microscope consisting of an objective lens of 20 cm focal length and an eyepiece of 10 cm focal length. When there are 45 cm between the lenses, the overall magnification of the fuzz balls is an admittedly pretty lame 2.0. Determine how far the fuzz balls are from the objective lens.

0.6 Sketchy Answers

(1a) 39°.

(2) 4.7 m.

(5) 49°.

(6a) 1.2.

(6b) 47°.

(7a) The set of answers you should get — not necessarily in order, and without the ± signs — is 7.5, 30, and 60 cm.

(8) The set of answers you should get for the first part — again, not necessarily in order, and without the ± signs — is 3.8, 6.0, 8.6, and 10 cm.

(11a) 5.6 cm.

(11b) 1.9 cm.

(12a) Keeping an extra digit, 2.14 to 2.50 cm.

(12b) 2.38 cm.

(12c) −2.00 diopters.

(12d) 2.00 diopters.

(12e) 40.0 cm.

(12f) 66.7 cm.

(14a) 61 cm.

(14b) 49 times smaller.

(14c) 9.9 m.

(14d) 1.4.

(16) 40 cm.

Chapter 1

Vectors

A vector has both a magnitude and a direction.[1] Not likely to make your list of the most interesting things you've ever encountered, but there you have it.

Anyway, a vector can be visualized simply as an arrow, with the orientation of the arrow corresponding to the vector's direction and the length of the arrow corresponding to the vector's magnitude. Many physical quantities are vectors; perhaps the simplest, most intuitive example of a vector is a spatial displacement: you move a certain distance in a certain direction, say, 100 m north.

Conventionally, symbols representing vectors are boldface in a books: for example, vector $\mathbf{V}$. Since it is hard to draw boldface letters, the convention in handwriting is to put a little arrow over any symbol representing a vector: $\vec{V}$. The magnitude (length) of the vector is denoted by plain type in books and in handwriting: V. The magnitude is sometimes also denoted by putting an absolute-value symbol around the vector symbol: $|\mathbf{V}|$ in books or $|\vec{V}|$ by hand.

If you think in terms of arrows representing spatial displacements, then the geometric definitions of vector addition, subtraction, and scalar multiplication are fairly intuitive: to add vectors, you simply place them "head to tail" one after another; the sum (also known as the resultant) is then the vector (the arrow) that goes from the tail of the first vector to the head of the last, as shown in fig. (1.1). In terms of spatial displacements, the sum of the vectors is your overall displacement — the displacement that would take you, as the crow flies, straight from your original to your final location. As you can easily convince yourself by sketching an example or two — or, if you are lazy, looking at fig. (1.2) — , vector addition is commutative: the

[1] As opposed to a scalar, which is simply a number. A scalar may be negative as well as positive, and it may have physical dimensions (units) associated with it, but it does not have a direction.

Also, note that technical terms are typeset in colored text when they first appear.

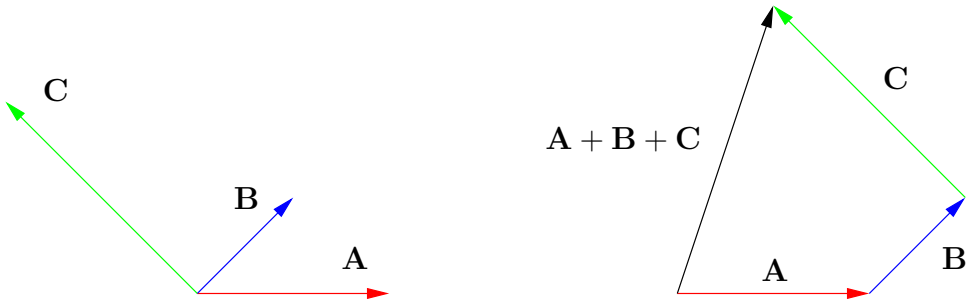


Figure 1.1: Vector Addition

order in which vectors are added makes no difference. As you can see from fig. (1.3), vector addition is also associative: adding $\mathbf{A}$ to $\mathbf{B}$ and then adding their sum to $\mathbf{c}$ is the same as adding $\mathbf{B}$ to $\mathbf{C}$ and then adding their sum to $\mathbf{A}$: $(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$.

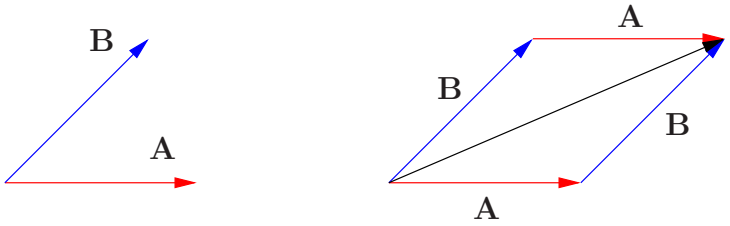


Figure 1.2: $\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$

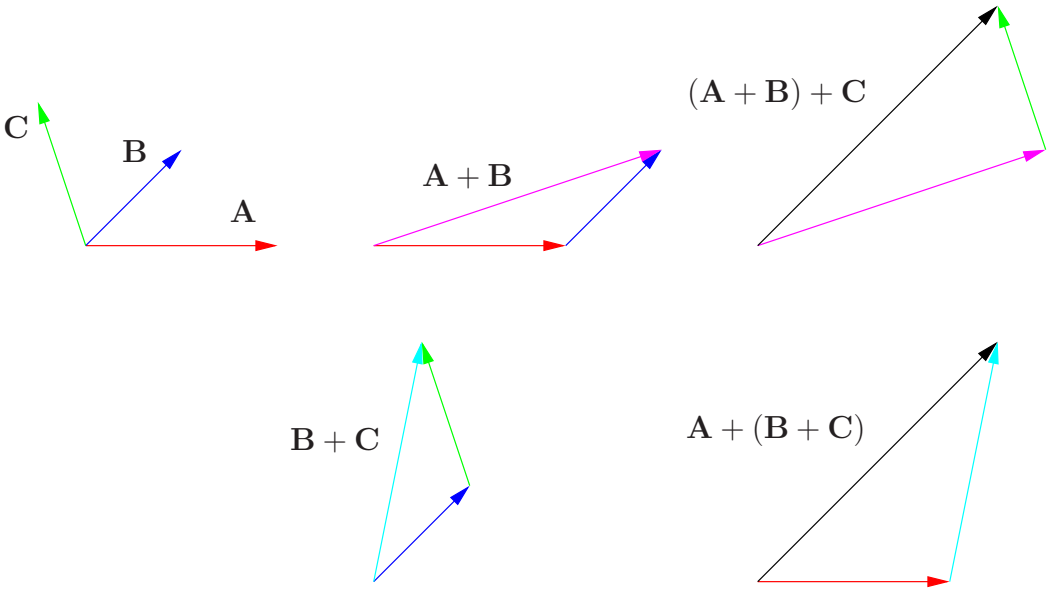


Figure 1.3: $(\mathbf{A}+\mathbf{B})+\mathbf{C}=\mathbf{A}+(\mathbf{B}+\mathbf{C})$

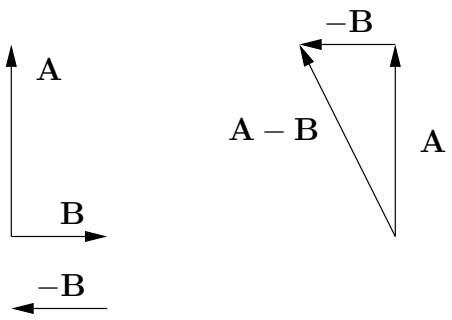


Figure 1.4: Vector Subtraction

The inverse $-\mathbf{V}$ of a vector $\mathbf{V}$ is the vector that "undoes" $\mathbf{V}$. That is, $-\mathbf{V}$ is the same as $\mathbf{V}$, just flipped around so that it points in the opposite direction, so that if $\mathbf{V}$ represented a displacement of 100 m to the east, then $-\mathbf{V}$ would represent a displacement of 100 m to the west. The most natural way to define subtraction is then simply addition of the inverse:

$$\mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B})$$

as shown in fig. (1.4).

Scalar multiplication means multiplying a vector by a number, that is, changing its length by some factor. For example, if $\mathbf{B}$ represents a displacement of 100 m to the east, then twice that displacement, $2\mathbf{B}$, would simply double the distance you travel in that same direction: 200 m to the east. You can also multiply vectors by fractional or negative numbers: $\frac{1}{2}\mathbf{B}$ would be 50 m to the east, and $-\frac{1}{2}\mathbf{B}$ would be 50 m to the *west*. Fig. (1.5) shows some multiples of a vector $\mathbf{B}$.

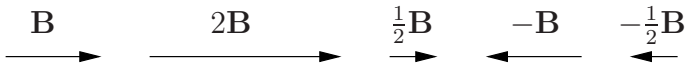


Figure 1.5: Scalar Multiplication

When you work with vectors, making rough sketches with arrows may help you visualize what you are doing with them. Graphical ("head-to-tail") addition is, however, too clumsy and inexact for doing calculations; instead, we will do vector operations like addition algebraically (that is, analytically) by means of components. The x and y components a_x and a_y of a vector $\mathbf{a}$ are the projections of $\mathbf{a}$ onto the x and y axes, as shown in fig. (1.6).

If θ is the conventional counterclockwise angle from the positive x axis,[2] then, as you can see from the diagram above, the trigonometry and Pytha-

[2]There is nothing magical about the counterclockwise angle from the positive x axis; its use is purely convention. As long as you make clear which angle you mean, you can specify the direction of a vector by giving any angle you want.

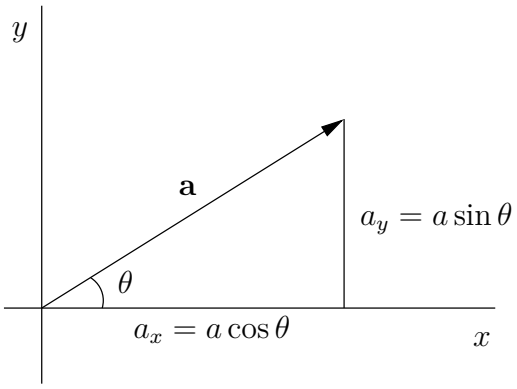


Figure 1.6: Components of a Vector

gorean theorem yield

$$a_x = a\cos\theta \qquad a_y = a\sin\theta$$
$$a = \sqrt{a_x^2 + a_y^2} \qquad \tan\theta = \frac{a_y}{a_x}$$

These pairs of relations, for a_x and a_y in terms of a and θ and for a and θ in terms of a_x and a_y, tell us that there are two different ways of specifying what vector you're talking about: either you can specify the magnitude a and the direction in terms of the angle θ, or you can specify the components a_x and a_y; these two sets of information are completely equivalent.[3]

When we add vectors analytically, therefore, *the component of the sum is the sum of the components.* Thus, as shown in fig. (1.7) — though just for the x components, so that the diagram doesn't get too cluttered — , if

$$\mathbf{c} = \mathbf{a} + \mathbf{b}$$

then

$$c_x = a_x + b_x$$
$$c_y = a_y + b_y$$

The general procedure for working with vector sums is consequently very straightforward:

1. Determine the x and y components of each vector that you are adding.
2. Add these x and y components to determine the x and y components of the resultant.

[3]Notice that it takes two pieces of information to specify our two-dimensional vector **a**. More generally, in n dimensions it takes n pieces of information to specify a vector: we can give the vector's n components along the coordinate axes, or we can give the vector's magnitude and, to specify its direction, $n-1$ angles. In three-dimensional spherical coordinates, for example, we need to specify two angles, the polar and azimuthal angles.

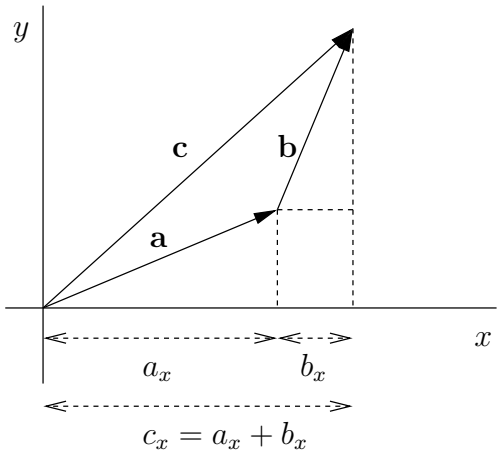


Figure 1.7: Adding Vectors by Components

3. If necessary (that is, if you need to give the magnitude of the resultant or its direction in terms of an angle), use the relations

$$a = \sqrt{a_x^2 + a_y^2} \qquad \text{and} \qquad \tan\theta = \frac{a_y}{a_x}$$

There is, however, one catch to be aware of: when solving $\tan\theta = a_y/a_x$, if you use your calculator to take the inverse tangent, it will tacitly give the *principal value* of the angle — which may or may not be what you want. If, for example, vectors **a** and **b** have components $(a_x, a_y) = (-1, \sqrt{3})$ and $(b_x, b_y) = (1, -\sqrt{3})$, your calculator will for both vectors give you $\theta = -30°$, which would correspond to the fourth quadrant. From its components, we see that vector **b** is indeed in the fourth quadrant, and $\theta = -30°$ is therefore the correct angle for it. Vector **a**, however, lies in the second quadrant, so that we want a different branch of the arctangent: not $\theta = -63.4°$, but $\theta = -30+180 = 150°$. The only issue is whether you want to add or subtract 180°, and if you make a quick sketch of the vector in question, this will be clear from the quadrant in which the vector lies.

1.1 Unit Vectors

A unit vector is simply a vector that has unit length (that is, a length of 1). To distinguish them from other vectors, unit vectors, both typeset and handwritten, are denoted by a "hat": $\hat{\mathbf{u}}$. Any vector **a** can be made into a

unit vector by dividing it by its own magnitude:[4]

$$\hat{\mathbf{a}} = \frac{\mathbf{a}}{a}$$

This $\hat{\mathbf{a}}$ is then indeed of unit length:

$$|\hat{\mathbf{a}}| = \left|\frac{\mathbf{a}}{a}\right| = \frac{|\mathbf{a}|}{a} = \frac{a}{a} = 1$$

Unit vectors allow us to express vectors in a way that is frequently convenient for calculations. If we define $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ to be unit vectors pointing in the positive x, positive y, and positive z directions,[5] respectively, then any three dimensional vector $\mathbf{a}$ can be written in terms of its components a_x, a_y, and a_z as

$$\mathbf{a} = a_x\hat{\mathbf{x}} + a_y\hat{\mathbf{y}} + a_z\hat{\mathbf{z}} \tag{1.1}$$

Strictly speaking, a_x, a_y, and a_z are known as the scalar components of $\mathbf{a}$ because a_x, a_y, and a_z are scalar quantities; $a_x\hat{\mathbf{x}}$, $a_y\hat{\mathbf{y}}$, and $a_z\hat{\mathbf{z}}$ are known as the vector components of $\mathbf{a}$ because each of them has, by virtue of the $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, or $\hat{\mathbf{z}}$, a direction as well as a magnitude and therefore constitutes a vector quantity. Eq. (1.1) is telling us that we can reproduce the vector $\mathbf{a}$ by going a distance a_x in the positive x direction,[6] then a distance a_y in the positive y direction, and then a distance a_z in the positive z direction, as is shown for the case of a two-dimensional vector in fig. (1.8).

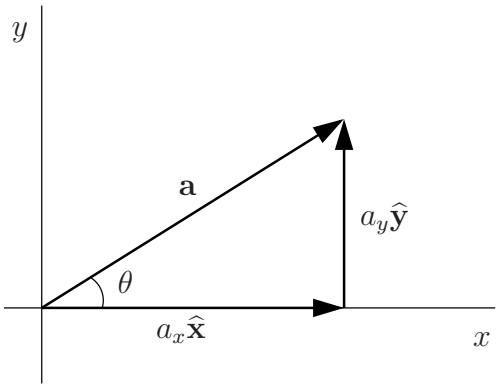


Figure 1.8: A Vector as the Sum of Its Vector Components

[4]Note that in doing so any dimensions associated with the vector $\mathbf{a}$ will cancel out. This is how it is possible to define a unit vector as having a length of 1 without specifying any physical units on the 1: no matter what the dimensions of $\mathbf{a}$, $\hat{\mathbf{a}}$ is dimensionless; the 1 is a pure number.

[5]Just so you know, some books use $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ instead of $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$.

[6]In the negative x direction, of course, if a_x is negative.

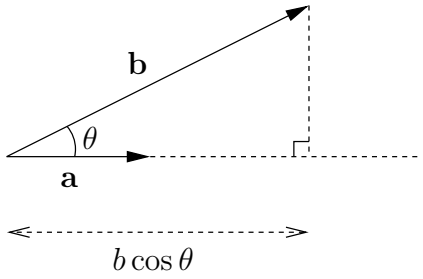


Figure 1.9: The Dot Product As a Projection

1.2 Dot & Cross Products

Two very useful binary vector operations are the dot product and the cross product.[7] Each of these has two equivalent definitions, one geometric and one analytic.

We will be making extensive use of dot and cross products. Eventually, that is — it will be a while before we have any occasion to apply them. For now, we will simply define them and introduce you to their properties.

1.2.1 The Dot Product

The dot product is so called because the notation for it is a dot between the two vectors: $\mathbf{a} \cdot \mathbf{b}$.[8] The geometric definition of the dot product is

$$\mathbf{a} \cdot \mathbf{b} = ab \cos\theta \tag{1.2}$$

where θ is the angle between the vectors $\mathbf{a}$ and $\mathbf{b}$. Note that the $b\cos\theta$ part of $ab\cos\theta$ can be regarded as the projection of $\mathbf{b}$ onto $\mathbf{a}$, as shown in fig. (1.9).[9]

The analytic definition of the dot product is[10]

$$\mathbf{a} \cdot \mathbf{b} = a_x b_x + a_y b_y + a_z b_z \tag{1.3}$$

To see that the geometric definition (1.2) of the dot product is equivalent to the analytic definition (1.3), consider a vector $\mathbf{c}$ that is the sum of two vectors $\mathbf{a}$ and $\mathbf{b}$. As shown in black in fig. (1.10), these three vectors form a triangle, where the angle Θ between $\mathbf{a}$ and $\mathbf{b}$ when they are added graphically head-to-tail is related to the angle θ of eq. (1.2) by

$$\Theta = \pi - \theta$$

[7] *Binary*, in case you're not familiar with the term, meaning involving *two* vectors.

[8] The dot product is also known as the scalar product because its value is a scalar, and, for reasons we won't go into, as the inner product.

[9] And by the symmetry of the definition, the $a\cos\theta$ part of $ab\cos\theta$ can of course likewise be regarded as the projection of $\mathbf{a}$ onto $\mathbf{b}$.

[10] For two-dimensional vectors, this of course reduces to simply $\mathbf{a} \cdot \mathbf{b} = a_x b_x + a_y b_y$.

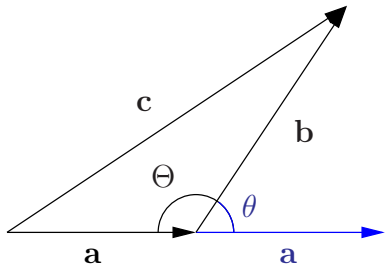


Figure 1.10: Monkeyshines with Dot Products

Evaluating the dot product of $\mathbf{c} = \mathbf{a} + \mathbf{b}$ with itself by the analytic definition (1.3) gives

$$\begin{aligned} \mathbf{c} \cdot \mathbf{c} &= (\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} + \mathbf{b}) \\ c_x c_x + c_y c_y + c_z c_z &= (\mathbf{a} + \mathbf{b})_x (\mathbf{a} + \mathbf{b})_x + (\mathbf{a} + \mathbf{b})_y (\mathbf{a} + \mathbf{b})_y \\ &\quad + (\mathbf{a} + \mathbf{b})_z (\mathbf{a} + \mathbf{b})_z \\ c_x^2 + c_y^2 + c_z^2 &= (a_x + b_x)^2 + (a_y + b_y)^2 + (a_z + b_z)^2 \end{aligned}$$

and hence, if we expand the squares on the right-hand side and regroup terms,

$$\begin{aligned} c_x^2 + c_y^2 + c_z^2 &= (a_x^2 + 2a_x b_x + b_x^2) + (a_y^2 + 2a_y b_y + b_y^2) \\ &\quad + (a_z^2 + 2a_z b_z + b_z^2) \\ &= (a_x^2 + a_y^2 + a_z^2) + (b_x^2 + b_y^2 + b_z^2) \\ &\quad + 2(a_x b_x + a_y b_y + a_z b_z) \end{aligned} \tag{1.4}$$

By the Pythagorean theorem, $c_x^2 + c_y^2 + c_z^2$ is just the squared magnitude c^2 of the vector $\mathbf{c}$, and likewise for the sums of the squares of the components of $\mathbf{a}$ and $\mathbf{b}$. Thus eq. (1.4) reduces to

$$c^2 = a^2 + b^2 + 2(a_x b_x + a_y b_y + a_z b_z) \tag{1.5}$$

If we now apply the law of cosines to the triangle of fig. (1.10), we also have

$$\begin{aligned} c^2 &= a^2 + b^2 - 2ab\cos\Theta \\ &= a^2 + b^2 - 2ab\cos(\pi - \theta) \\ &= a^2 + b^2 + 2ab\cos\theta \end{aligned} \tag{1.6}$$

Comparing eqq. (1.5) and (1.6), we see that

$$a_x b_x + a_y b_y + a_z b_z = ab\cos\theta$$

Since this holds for any two vectors $\mathbf{a}$ and $\mathbf{b}$, the analytic and geometric definitions (1.3) and (1.2) of the dot product are therefore equivalent.

Note that the dot product is independent of our choice of coordinate system, that is, of how we orient our xyz axes. This independence is obvious for the geometric dot product: $\mathbf{a}\cdot\mathbf{b} = ab\cos\theta$ depends only on the magnitudes of the vectors $\mathbf{a}$ and $\mathbf{b}$ and on the angle between them, which have nothing to do with how we set up our coordinate axes. If we didn't know any better, it might seem that the analytic dot product would depend on the choice of coordinate system — in different coordinate systems the components (a_x, a_y, a_z) and (b_x, b_y, b_z) of the vectors $\mathbf{a}$ and $\mathbf{b}$ will differ — , but appearances can be deceiving: since we have just established the equivalence of the analytic and geometric dot products, the analytic dot product must also be independent of the choice of coordinate system.[11]

As you can see from both the geometric and analytic definitions, the dot product is symmetric and therefore commutative: the order of the vectors in the dot product doesn't matter. Also note that since the result of taking the dot product is a scalar, not a vector, constructs like $\mathbf{a}\cdot(\mathbf{b}\cdot\mathbf{c})$ are complete nonsense.

General properties of the dot product, which can be shown from either the geometric or analytic definition, and some of which will be addressed in the homework problems, include

$$\begin{aligned}
\mathbf{a}\cdot\mathbf{b} &= \mathbf{b}\cdot\mathbf{a} \\
\mathbf{a}\cdot(\mathbf{b}+\mathbf{c}) &= \mathbf{a}\cdot\mathbf{b} + \mathbf{a}\cdot\mathbf{c} \\
\mathbf{a}\cdot\mathbf{a} &= a^2 \\
\mathbf{a}\cdot\mathbf{b} = 0 \;\Rightarrow\; \mathbf{a}\perp\mathbf{b} &\qquad (\text{or } a=0 \text{ or } b=0)
\end{aligned}$$

1.2.2 The Cross Product

The cross product is so called because the notation for it is, as would probably be your first guess, a cross between the two vectors: $\mathbf{a}\times\mathbf{b}$.[12] The geometric

[11] The big-people way to think of the dot product is in terms of vectors and rotation operators in linear algebra: the dot product would be represented by the product of a row vector $\mathbf{a}^t$ on the left with a column vector $\mathbf{b}$ on the right, where the superscript t indicates the transpose. The effect of a rotation on the vectors $\mathbf{a}$ and $\mathbf{b}$ can be represented by a rotation matrix $\mathcal{O}$ that is an element, in n dimensions, of $O(n)$ (the orthogonal group of order n): as a result of the rotation, $\mathbf{a}\to\mathcal{O}\mathbf{a}$ and $\mathbf{b}\to\mathcal{O}\mathbf{b}$. Since for $O(n)$ the transpose is the inverse ($\mathcal{O}^t = \mathcal{O}^{-1}$),

$$\mathbf{a}^t\mathbf{b} \to (\mathcal{O}\mathbf{a})^t\mathcal{O}\mathbf{b} = \mathbf{a}^t\mathcal{O}^t\mathcal{O}\mathbf{b} = \mathbf{a}^t\mathcal{O}^{-1}\mathcal{O}\mathbf{b} = \mathbf{a}^t\mathbf{b}$$

In fact, the invariance of $\mathbf{a}\cdot\mathbf{b}$ can be used to define the rotation group $O(n)$.

[12] The cross product is also known as the vector product because its value is a vector and, for reasons we won't go into, as the outer product.

Although strictly speaking the value of the cross product is not a true vector. When all three spatial axes are inverted, so that $(x, y, z) \to (-x, -y, -z)$, then true vectors,

definition of the cross product is

$$\mathbf{a} \times \mathbf{b} = ab \sin\theta \, \hat{\mathbf{u}} \tag{1.7}$$

where θ is the angle between the vectors $\mathbf{a}$ and $\mathbf{b}$ and $\hat{\mathbf{u}}$ is perpendicular to the plane of $\mathbf{a}$ and $\mathbf{b}$ *in a right-handed sense*. "In a right-handed sense" means that you apply the right-hand rule to the vectors $\mathbf{a}$ and $\mathbf{b}$: you take your right hand,[13] with fingers straight out and thumb in the same plane but extended at a right angle, and place it along the first vector in the cross product ($\mathbf{a}$) with your palm facing toward the second vector in the cross product ($\mathbf{b}$), so that you can curl your fingers toward that second vector ($\mathbf{b}$).[14] The direction in which your thumb points is the direction of $\mathbf{a} \times \mathbf{b}$.

This right-handedness of the cross product is necessary to avoid ambiguity: for any two vectors $\mathbf{a}$ and $\mathbf{b}$,[15] there are two directions perpendicular to the plane formed by $\mathbf{a}$ and $\mathbf{b}$. The vectors shown in fig. (1.9) on p.71, for example, lie in the plane of the page, so that both into the page and out of the page would qualify as being perpendicular to the plane that they form. Applying the right-hand rule to the vectors shown in fig. (1.9) leaves your thumb pointing out of the page, however, so that $\mathbf{a} \times \mathbf{b}$ is unambiguously out of the page.

Cross products can, of course, end up being in any direction, but for many of the vectors we deal with, they will end up being either into or out of the page. The conventional notation for the directions "into the page" and "out of the page" is $\otimes$ and $\odot$, respectively.

The equivalent analytic definition of the cross product is

$$\mathbf{a} \times \mathbf{b} = (a_y b_z - a_z b_y)\,\hat{\mathbf{x}} - (a_x b_z - a_z b_x)\,\hat{\mathbf{y}} + (a_x b_y - a_y b_x)\,\hat{\mathbf{z}} \tag{1.8}$$

Because the cross product is a vector[16] and not a scalar, the proof of the equivalence of its analytic and geometric definitions is a bit more involved than the equivalence proof for the dot product: we need to show that both the magnitudes and the directions given by the analytic and geometric definitions

---

such as $\mathbf{a}$ and $\mathbf{b}$, are also inverted: $\mathbf{a} \to -\mathbf{a}$ and $\mathbf{b} \to -\mathbf{b}$. But under inversion of the spatial axes the cross product will *not* change sign: $\mathbf{a} \times \mathbf{b} \to (-\mathbf{a}) \times (-\mathbf{b}) = \mathbf{a} \times \mathbf{b}$. (This can also be seen from the analytic definition, since $\mathbf{a} \to -\mathbf{a}$ and $\mathbf{b} \to -\mathbf{b}$ means that $(a_x, a_y, a_z) \to (-a_x, -a_y, -a_z)$ and $(b_x, b_y, b_z) \to (-b_x, -b_y, -b_z)$.) A vector that, like a cross product, does not change sign under spatial inversion is called an axial vector.

[13] This part of the prescription is of critical importance. There is a strong tendency to use your free hand when applying the right-hand rule — which, if you write with your right hand, is of course not what you want.

[14] If you can curl your fingers toward $\mathbf{b}$ with your palm facing away from it, it will make an amazing class demonstration.

[15] Any two linearly independent vectors $\mathbf{a}$ and $\mathbf{b}$, that is — for those of you into pedantry.

[16] Again, the cross product actually yields an axial vector, but this doesn't affect our argument.

agree. We will do this in three steps by showing that the result of the analytic cross product agrees in magnitude with that of the geometric cross product, that it is perpendicular to the plane of the vectors being crossed, and that it is perpendicular in a right-handed sense.

First, the magnitudes: As we saw in our proof of the equivalence of the definitions of the dot product in the preceding section, the dot product of any vector with itself yields the square of its magnitude. If we apply this to the cross product $\mathbf{a} \times \mathbf{b}$, we have

$$\begin{aligned} |\mathbf{a} \times \mathbf{b}|^2 &= (\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{a} \times \mathbf{b}) \\ &= (\mathbf{a} \times \mathbf{b})_x (\mathbf{a} \times \mathbf{b})_x + (\mathbf{a} \times \mathbf{b})_y (\mathbf{a} \times \mathbf{b})_y \\ &\qquad + (\mathbf{a} \times \mathbf{b})_z (\mathbf{a} \times \mathbf{b})_z \end{aligned}$$

which, if we substitute the components of $\mathbf{a} \times \mathbf{b}$ from the analytic definition (1.8), expand the squares, and then very patiently regroup the terms, becomes

$$\begin{aligned} |\mathbf{a} \times \mathbf{b}|^2 &= (a_y b_z - a_z b_y)^2 + \Big(- (a_x b_z - a_z b_x) \Big)^2 \\ &\qquad + (a_x b_y - a_y b_x)^2 \\ &= (a_y^2 b_z^2 - 2 a_y a_z b_y b_z + a_z^2 b_y^2) + (a_x^2 b_z^2 - 2 a_x a_z b_x b_z + a_z^2 b_x^2) \\ &\qquad + (a_x^2 b_y^2 - 2 a_x a_y b_x b_y + a_y^2 b_x^2) \\ &= a_x^2 (b_y^2 + b_z^2) + a_y^2 (b_x^2 + b_z^2) + a_z^2 (b_x^2 + b_y^2) \\ &\qquad - 2(a_x a_y b_x b_y + a_y a_z b_y b_z + a_x a_z b_x b_z) \\ &= a_x^2 (b_y^2 + b_z^2) + a_y^2 (b_x^2 + b_z^2) + a_z^2 (b_x^2 + b_y^2) + (a_x^2 b_x^2 + a_y^2 b_y^2 + a_z^2 b_z^2) \\ &\qquad - (a_x^2 b_x^2 + a_y^2 b_y^2 + a_z^2 b_z^2) - 2(a_x a_y b_x b_y + a_y a_z b_y b_z + a_x a_z b_x b_z) \\ &= a_x^2 (b_x^2 + b_y^2 + b_z^2) + a_y^2 (b_x^2 + b_y^2 + b_z^2) + a_z^2 (b_x^2 + b_y^2 + b_z^2) \\ &\qquad - (a_x^2 b_x^2 + a_y^2 b_y^2 + a_z^2 b_z^2 + 2 a_x a_y b_x b_y + 2 a_y a_z b_y b_z + 2 a_x a_z b_x b_z) \\ &= (a_x^2 + a_y^2 + a_z^2)(b_x^2 + b_y^2 + b_z^2) \\ &\qquad - (a_x b_x + a_y b_y + a_z b_z)^2 \\ &= a^2 b^2 - (\mathbf{a} \cdot \mathbf{b})^2 \end{aligned}$$

Having already established the equivalence of the analytic and geometric definitions of the dot product, we can use $\mathbf{a} \cdot \mathbf{b} = ab \cos\theta$ to rewrite this as

$$\begin{aligned} |\mathbf{a} \times \mathbf{b}|^2 &= a^2 b^2 - (ab \cos\theta)^2 \\ &= a^2 b^2 (1 - \cos^2\theta) \\ &= a^2 b^2 \sin^2\theta \end{aligned}$$

Since magnitudes of vectors are always positive, and likewise the value of $\sin\theta$ for $0 \leq \theta \leq \pi$, when we take the square root we want the positive roots

on both sides:

$$|\mathbf{a} \times \mathbf{b}| = ab \sin\theta$$

which is exactly what the geometric cross product would give us. The geometric and analytic definitions of the cross product thus give the same result for the magnitude of $\mathbf{a} \times \mathbf{b}$.

The next step is to show that the result of the analytic cross product is perpendicular to the plane of the vectors being crossed. This is most easily demonstrated by noting that if $\mathbf{a}$ and $\mathbf{c}$ are nonzero vectors, then $\mathbf{a} \cdot \mathbf{c} = 0$ means that $\mathbf{a}$ is perpendicular to $\mathbf{c}$: if we denote the angle between $\mathbf{a}$ and $\mathbf{c}$ by θ_{ac}, then when a and c are nonzero we can divide

$$0 = \mathbf{a} \cdot \mathbf{c} = ac \cos\theta_{ac}$$

through by ac to obtain $\cos\theta_{ac} = 0$ and hence $\theta_{ac} = \frac{\pi}{2}$. In particular, when $\mathbf{c} = \mathbf{a} \times \mathbf{b}$, substituting the components of $\mathbf{a} \times \mathbf{b}$ from the analytic definition (1.8) gives

$$\begin{aligned}
\mathbf{a} \cdot (\mathbf{a} \times \mathbf{b}) &= a_x(\mathbf{a} \times \mathbf{b})_x + a_y(\mathbf{a} \times \mathbf{b})_y + a_z(\mathbf{a} \times \mathbf{b})_z \\
&= a_x(a_y b_z - a_z b_y) + a_y\Big(-(a_x b_z - a_z b_x)\Big) + a_z(a_x b_y - a_y b_x) \\
&= (a_x a_y - a_y a_x)b_z + (-a_x a_z + a_z a_x)b_y + (a_y a_z - a_z a_y)b_x \\
&= 0
\end{aligned}$$

And $\mathbf{b} \cdot (\mathbf{a} \times \mathbf{b})$ would similarly work out to zero. The result of the analytic cross product is therefore a vector perpendicular to both $\mathbf{a}$ and $\mathbf{b}$, that is, perpendicular to the plane of $\mathbf{a}$ and $\mathbf{b}$.

The final step in establishing the equivalence of the analytic and geometric definitions of the cross product is to show that the analytic result for $\mathbf{a} \times \mathbf{b}$ is not only perpendicular to the plane of $\mathbf{a}$ and $\mathbf{b}$, but perpendicular in a right-handed sense. To see this we can set up a coordinate system in which the x axis lies along $\mathbf{a}$ and $\mathbf{b}$ is in the upper xy plane, as shown in fig. (1.11). Then

$$\begin{aligned}
a_x &= a & a_y &= 0 & a_z &= 0 \\
b_x &= b\cos\theta & b_y &= b\sin\theta & b_z &= 0
\end{aligned}$$

and the analytic definition (1.8) of the cross product gives

$$\begin{aligned}
\mathbf{a} \times \mathbf{b} &= (a_y b_z - a_z b_y)\,\hat{\mathbf{x}} - (a_x b_z - a_z b_x)\,\hat{\mathbf{y}} + (a_x b_y - a_y b_x)\,\hat{\mathbf{z}} \\
&= ab\sin\theta\,\hat{\mathbf{z}}
\end{aligned}$$

Since, as noted above, $\sin\theta$ and the magnitudes a and b are all positive, the analytic result for $\mathbf{a} \times \mathbf{b}$ is in the positive z direction, which is indeed in a

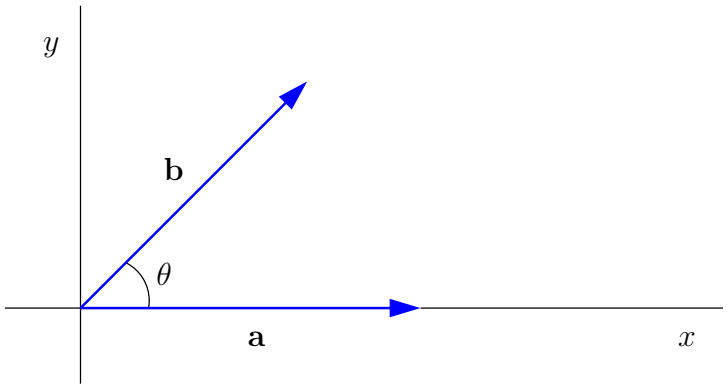


Figure 1.11: Monkeyshines with Cross Products

right-handed sense relative to **a** and **b**. And this, we are sure you will be glad to hear, completes the proof of the equivalence of the analytic and geometric definitions of the cross product. We can now get on with our lives.

Those of you who have some familiarity with determinants will find that the analytic definition (1.8) of the cross product can be much more easily expressed and remembered as

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{vmatrix}$$

Those of you with no familiarity with determinants are pretty much screwed.[17]

Anyway, for two-dimensional vectors in the xy plane, the analytic cross product reduces to

$$\mathbf{a} \times \mathbf{b} = (a_x b_y - a_y b_x)\,\hat{\mathbf{z}}$$

Note that, as you can see from interchanging a and b in the analytic definition (1.8), the cross product is anticommutative, that is,

$$\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}$$

In the geometric definition, interchanging a and b would mean you curl your fingers from **b** to **a** instead of from **a** to **b**, which will leave your thumb and

[17] Actually, it isn't that difficult to reproduce eq. (1.8): note that each of the six terms on the right-hand side of eq. (1.8) consists of the product of a component of **a**, and component of **b**, and a unit vector, with each of those three animals involving a different member of the set $\{x, y, z\}$. The positive terms in eq. (1.8) are the ones corresponding the cyclic orderings of $\{x, y, z\}$ obtained by cycling the front guy to the back: xyz, yzx, and zxy. The negative terms in eq. (1.8) are the ones corresponding the anticyclic orderings of $\{x, y, z\}$ that also require a flip (an interchange) of two guys: yxz, zyx, and xzy. So you can just set up the three cyclic contributions to $\mathbf{a} \times \mathbf{b}$, and the remaining terms, which have the opposite sign, will just be those with the components of **a** and **b** flipped.

therefore the cross product pointing in the opposite direction. Thus changing the order of the vectors in the cross product reverses its direction — which, for a vector, is equivalent to multiplying it by -1.

General properties of the cross product, which can be shown from either the geometric or analytic definition, and some of which will be addressed in the homework problems, include

$$\begin{aligned}
\mathbf{a}\times\mathbf{b} &= -\mathbf{b}\times\mathbf{a} \\
\mathbf{a}\times(\mathbf{b}+\mathbf{c}) &= \mathbf{a}\times\mathbf{b}+\mathbf{a}\times\mathbf{c} \\
\mathbf{a}\times\mathbf{a} &= 0 \\
\mathbf{a}\times(\mathbf{b}\times\mathbf{c}) &= \mathbf{a}\cdot\mathbf{c}\,\mathbf{b}-\mathbf{a}\cdot\mathbf{b}\,\mathbf{c} \\
\mathbf{a}\times\mathbf{b}=0 &\;\Rightarrow\; \mathbf{a}\parallel\mathbf{b} \qquad (\text{or } a=0 \text{ or } b=0)
\end{aligned}$$

where $\mathbf{a}\parallel\mathbf{b}$ means that $\mathbf{a}$ is either parallel or antiparallel to $\mathbf{b}$.

1.2.3 Some Special Cases

Just how dot and cross products will be useful to us will remain a mystery for now. We will, however, note the following very useful relations for dot and cross products involving the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$:

- Any of the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ dotted with itself yields unity (*duh!*):

$$\hat{\mathbf{x}}\cdot\hat{\mathbf{x}}=1 \qquad \hat{\mathbf{y}}\cdot\hat{\mathbf{y}}=1 \qquad \hat{\mathbf{z}}\cdot\hat{\mathbf{z}}=1$$

- Because $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are all perpendicular to each other, their dot products with each other all vanish:[18]

$$\hat{\mathbf{x}}\cdot\hat{\mathbf{y}}=\hat{\mathbf{y}}\cdot\hat{\mathbf{x}}=0 \qquad \hat{\mathbf{x}}\cdot\hat{\mathbf{z}}=\hat{\mathbf{z}}\cdot\hat{\mathbf{x}}=0 \qquad \hat{\mathbf{y}}\cdot\hat{\mathbf{z}}=\hat{\mathbf{z}}\cdot\hat{\mathbf{y}}=0$$

- As for any vectors, the cross product of any of the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ with itself vanishes:

$$\hat{\mathbf{x}}\times\hat{\mathbf{x}}=0 \qquad \hat{\mathbf{y}}\times\hat{\mathbf{y}}=0 \qquad \hat{\mathbf{z}}\times\hat{\mathbf{z}}=0$$

- And, as you should be able to work out using the right-hand rule,

$$\begin{aligned}
&\hat{\mathbf{x}}\times\hat{\mathbf{y}}=\hat{\mathbf{z}} \qquad &&\hat{\mathbf{y}}\times\hat{\mathbf{z}}=\hat{\mathbf{x}} \qquad &&\hat{\mathbf{z}}\times\hat{\mathbf{x}}=\hat{\mathbf{y}} \\
&\hat{\mathbf{y}}\times\hat{\mathbf{x}}=-\hat{\mathbf{z}} \qquad &&\hat{\mathbf{z}}\times\hat{\mathbf{y}}=-\hat{\mathbf{x}} \qquad &&\hat{\mathbf{x}}\times\hat{\mathbf{z}}=-\hat{\mathbf{y}}
\end{aligned}$$

[18] *Vanish*, if you haven't seen this usage before, is the big-people way to say *equal zero*.

1.3 Problems

1. (Warning: This problem is excruciatingly boring.)

   (a) Some stupid displacement vector has a magnitude of 5.0 m and makes a counterclockwise angle of 37° with the positive x axis. Determine the x and y components of the vector.

   (b) Another stupid vector has a magnitude of 3.0 m and makes a counterclockwise angle of $-57°$ with the positive x axis. Determine the x and y components of the vector.

   (c) Still another stupid vector has x and y components of 2.5 m and 5.0 m, respectively. Determine the magnitude of this vector and the counterclockwise angle it makes with the positive x axis.

   (d) Yet another stupid vector — long is the night that never finds the day — has x and y components of -2.5 m and -5.0 m, respectively. Determine the magnitude of this vector and the counterclockwise angle it makes with the positive x axis.

   (e) Using graph paper, a ruler, and a protractor (or a drawing program, if you want), add the vectors of #1a and #1b *graphically*, to determine the magnitude of the resultant.

   (f) Take the *difference* of the vectors of #1a and #1b *analytically*, to determine the magnitude of the difference.

2. After being severely beaten about the head with a blunt object, Barney stumbles 4.0 m north, then 2.0 m southwest, then 3.0 m east. Using graph paper and a ruler (or a drawing program, if you want), draw Barney's path. *Graphically* determine Barney's net displacement by measuring directly from your graph with the ruler. Determine the x and y (east-west and north-south) components of Barney's net displacement in the same way.

3. Repeat #2, but this time do the calculations *analytically*.

4. Three vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ are related to each other by $\mathbf{a} + \mathbf{b} = \mathbf{c}$. The vector $\mathbf{a}$ has magnitude 2.0 and makes a 135° angle with the positive x axis. The vector $\mathbf{c}$ has x and y components of 3.0 and -1.5, respectively. Determine the magnitude and direction of the vector $\mathbf{b}$.

5. The same vectors $\mathbf{a}$ and $\mathbf{c}$ — as before, the vector $\mathbf{a}$ has magnitude 2.0 and makes a 135° angle with the positive x axis, and the vector $\mathbf{c}$ has x and y components of 3.0 and -1.5 — are now related by $3\mathbf{a} - 2\mathbf{b} + \frac{1}{2}\mathbf{c} = 0$. Determine the magnitude and direction of the vector $\mathbf{b}$.

6. (a) Solve
$$\alpha\mathbf{a} - \beta\mathbf{b} = \gamma\mathbf{c} \tag{1.9}$$
for $\mathbf{b}$ in terms of α, β, γ, $\mathbf{a}$ and $\mathbf{c}$.

   (b) Formally, what are the x, y, and z components of $\mathbf{b}$? (There is a large *Duh!* factor in this part.)

   (c) Formally, what are the x, y, and z components of eq. (1.9). (There is a large *Duh!* factor in this part, too.)

   (d) Formally determine the component of $\mathbf{b}$ along the direction of a vector $\mathbf{u}$, in terms of α, β, γ, $\mathbf{a}$, $\mathbf{c}$, and $\mathbf{u}$.

   (e) Similarly, what is the component of eq. (1.9) along the direction of $\mathbf{u}$?

   (f) How would you go about solving eq. (1.9) for β in terms of α, γ, $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$? Do it.

7. Vector $\mathbf{a}$ has a magnitude of 4 and makes a counterclockwise angle of $\frac{\pi}{6}$ with the positive x axis. Vector $\mathbf{b}$ has a magnitude of 2 and makes a counterclockwise angle of $-\frac{\pi}{6}$ with the positive x axis.

   (a) Use the geometric definition to evaluate $\mathbf{a} \cdot \mathbf{b}$.

   (b) Show that you get the same result using the analytic definition of the dot product.

   (c) Use the geometric definition to evaluate $\mathbf{a} \times \mathbf{b}$.

   (d) Show that you get the same result using the analytic definition of the cross product.

8. If $\mathbf{a} = \sqrt{3}\,\hat{\mathbf{x}} + \hat{\mathbf{y}}$ and $\mathbf{b} = \sqrt{3}\,\hat{\mathbf{x}} - 3\hat{\mathbf{y}}$,

   (a) Determine $\mathbf{a} \cdot \mathbf{b}$.

   (b) Determine $\mathbf{a} \times \mathbf{b}$.

   (c) Determine $\widehat{\mathbf{a}}$ and $\widehat{\mathbf{b}}$.

   (d) Determine the angle that $\mathbf{a}$ makes with $\mathbf{a} - \mathbf{b}$.

   (e) Determine the angle that $\mathbf{a}$ makes with $\mathbf{a} \times \mathbf{b}$.

9. If $\mathbf{a} = \hat{\mathbf{x}} + \hat{\mathbf{y}}$, $\mathbf{b} = 3\hat{\mathbf{x}} + 4\hat{\mathbf{y}}$, and $\mathbf{c} = -4\hat{\mathbf{x}} + 3\hat{\mathbf{y}}$,

   (a) Determine $\mathbf{a} \cdot (\mathbf{b} + \mathbf{c})$.

   (b) Determine $\mathbf{a} \times (\mathbf{b} - \mathbf{c})$.

   (c) Determine $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$.

   (d) Determine $\mathbf{a} \times (\mathbf{b} \times \mathbf{c})$.

10. Show that if $\mathbf{a} \cdot \mathbf{b} = 0$ for two nonzero vectors $\mathbf{a}$ and $\mathbf{b}$, then $\mathbf{a}$ and $\mathbf{b}$ are perpendicular to each other.

11. Show that if $\mathbf{a} \times \mathbf{b} = 0$ for two nonzero vectors $\mathbf{a}$ and $\mathbf{b}$, then $\mathbf{a}$ and $\mathbf{b}$ are either parallel or antiparallel to each other.

12. Show that for any vector $\mathbf{a}$, $\mathbf{a} \cdot \mathbf{a} = a^2$.

13. A cube with one corner at the origin and edges aligned with the x, y, and z axes has sides of length ℓ. Show that the angle between the long diagonal and any edge is $\cos^{-1}(1/\sqrt{3})$. See the footnote if you need a hint.[19]

14. Show that if $|\mathbf{a} + \mathbf{b}|^2 = a^2 + b^2$ for two nonzero vectors $\mathbf{a}$ and $\mathbf{b}$, then $\mathbf{a}$ and $\mathbf{b}$ are perpendicular to each other. Look back at #10 and #12 if you need a hint.

15. What does $|\mathbf{a} + \mathbf{b}|^2 = (a \pm b)^2$, where $\mathbf{a}$ and $\mathbf{b}$ are two nonzero vectors, tell you about $\mathbf{a}$ and $\mathbf{b}$? Look back at #14 if you need a hint.

16. Show that if $\mathbf{a} + \mathbf{b}$ is perpendicular to $\mathbf{a} - \mathbf{b}$ for two nonzero vectors $\mathbf{a}$ and $\mathbf{b}$, then $a = b$. Look back at #10 and #12 if you need a hint.

17. (a) Does $\mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \mathbf{c}$ always imply that $\mathbf{b} = \mathbf{c}$? If not, under what circumstances (that is, under what additional conditions) will this be the case?

    (b) Does $\mathbf{a} \times \mathbf{b} = \mathbf{a} \times \mathbf{c}$ always imply that $\mathbf{b} = \mathbf{c}$? If not, under what circumstances (that is, under what additional conditions) will this be the case?

18. Vector $\mathbf{a}$ has components $a_x = 3$, $a_y = 4$. Determine the components of a vector $\mathbf{b}$ that is perpendicular to $\mathbf{a}$ and has unit length. You should find that there are two such vectors $\mathbf{b}$.

19. Resolve the vector $\mathbf{b} = 3\hat{\mathbf{x}} + 4\hat{\mathbf{y}}$ into vector components $\mathbf{b}_\perp$ and $\mathbf{b}_\parallel$ that are, respectively, perpendicular and parallel to the vector $\mathbf{a} = \hat{\mathbf{x}} + \hat{\mathbf{y}}$.

20. If the edges of a parallelogram are given by $\mathbf{a}$ and $\mathbf{b}$, show that the area of the parallelogram is given by $|\mathbf{a} \times \mathbf{b}|$.

21. If the edges of a parallelepiped are given, starting from a common corner, by $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$, show that the volume of the parallelepiped is given by $\mathbf{a} \cdot \mathbf{b} \times \mathbf{c}$.

22. Three vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ satisfy the relations

$$\mathbf{a} + \mathbf{b} + \mathbf{c} = 0 \qquad \mathbf{a} \cdot \mathbf{c} = 0 \qquad \mathbf{b} \times \mathbf{c} = 0$$

What can you conclude about these vectors? Be sure to prove your assertions.

[19] Express the long diagonal as a vector in the form $\mathbf{d} = d_x\hat{\mathbf{x}} + d_y\hat{\mathbf{y}} + d_z\hat{\mathbf{z}}$.

23. Which of the following are sensible? Why or why not?

$\alpha + \beta\mathbf{a}$	$\alpha\mathbf{a} \cdot \mathbf{b}$	$\mathbf{a} + \mathbf{b} \cdot \mathbf{c}$	$e^{\mathbf{a}}$
$\alpha\beta\mathbf{a}$	$\alpha\mathbf{a} \times \mathbf{b}$	$\mathbf{a} + \mathbf{b} \times \mathbf{c}$	$e^{\mathbf{a}\cdot\mathbf{b}}$
$\alpha\mathbf{ab}$	$\mathbf{a} \cdot \mathbf{bc}$	$\sqrt{\mathbf{a}}$	$e^{\mathbf{a}\times\mathbf{b}}$
$\mathbf{a}^2$	$\mathbf{a} \times \mathbf{bc}$	$\sqrt{\mathbf{a} \cdot \mathbf{b}}$	$e^{\mathbf{a}^2}$

1.4 Sketchy Answers

(1a) 4.0 m, 3.0 m.

(1b) 1.6 m, −2.5 m.

(1c) 5.6 m at 63°.

(1d) 5.59 m at 243°.

(1e) You should get about 5.6. Roughly. More or less.

(1f) 6.0.

(2) You should get about 3.0 m, 1.6 m, 2.6 m.

(3) You should get 3.0 m, 1.6 m, 2.6 m.

(4) 5.3, −33°.

(5) 2.2, 128°.

(6a) $\mathbf{b} = \frac{1}{\beta}(\alpha\mathbf{a} - \gamma\mathbf{c})$.

(6b) $b_x = \frac{1}{\beta}(\alpha a_x - \gamma c_x)$, etc. Don't you feel stupid now? Or maybe angry.

(6c) $\alpha a_x - \beta b_x = \gamma c_x$, etc.

(6d) $b_u = \frac{1}{\beta}(\alpha\mathbf{a} - \gamma\mathbf{c}) \cdot \hat{\mathbf{u}}$.

(6e) $\alpha\mathbf{a} \cdot \hat{\mathbf{u}} - \beta\mathbf{b} \cdot \hat{\mathbf{u}} = \gamma\mathbf{c} \cdot \hat{\mathbf{u}}$.

(6f) $\beta = \frac{1}{b^2}\, \mathbf{b} \cdot (\alpha\mathbf{a} - \gamma\mathbf{c})$.

(7a) 4.

(7c) $-4\sqrt{3}\,\hat{\mathbf{z}}$.

(8a) 0.

(8b) $-4\sqrt{3}\,\hat{\mathbf{z}}$

(8c) $\frac{1}{2}(\sqrt{3}\,\hat{\mathbf{x}} + \hat{\mathbf{y}})$ and $\frac{1}{2}(\hat{\mathbf{x}} - \sqrt{3}\,\hat{\mathbf{y}})$.

(8d) $\frac{\pi}{3}$.

(8e) $\frac{\pi}{2}$. *Duh!*

(9a) 6.

(9b) $-6\hat{\mathbf{z}}$.

(9c) 0. You weren't really expecting otherwise, were you?

(9d) $25(\hat{\mathbf{x}} - \hat{\mathbf{y}})$.

(18) $\mathbf{b} = \pm\frac{4}{5}\hat{\mathbf{x}} \mp \frac{3}{5}\hat{\mathbf{y}}$.

(19) $\mathbf{b}_\perp = -\frac{1}{2}\hat{\mathbf{x}} + \frac{1}{2}\hat{\mathbf{y}}$, $\mathbf{b}_\parallel = \frac{7}{2}(\hat{\mathbf{x}} + \hat{\mathbf{y}})$.

(23) 9 are sensible, 7 are not. At least as far as you're concerned.

Chapter 2

Vector Calculus

2.1 Line Elements & Integrations

To specify location quantitatively in three-dimensional space, we could set up an xyz coordinate system; the values of the three coordinates x, y, and z would then tell us where we were along each of the three spatial directions. Location is a therefore a vector quantity, in the sense that we can draw a vector from the origin of our coordinate system to the point (x, y, z) where we are: it is not enough to say just how far we are from the origin; we must also specify the direction. The vector that goes from the origin to our location is called the position vector and is conventionally denoted by $\mathbf{r}$.[1] Since $\mathbf{r}$ extends from the origin to the point (x, y, z), the components of $\mathbf{r}$ are simply $r_x = x$, $r_y = y$, and $r_z = z$, as shown for the two-dimensional case in fig. (2.1):

$$\mathbf{r} = x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}$$

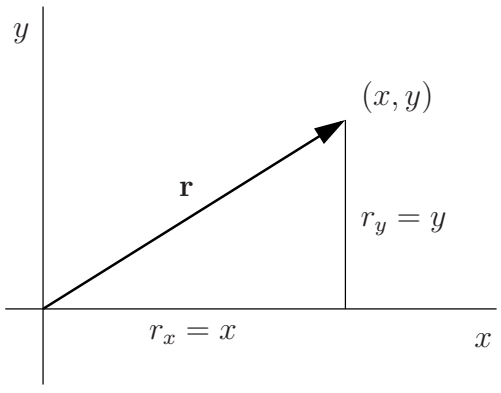


Figure 2.1: A Two-Dimensional Position Vector

[1] This notation is used because $\mathbf{r}$, since it starts from the origin, could also be thought of as a radius.

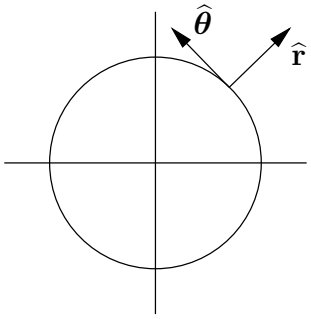


Figure 2.2: The Directions of $\widehat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$

The line element $d\mathbf{r}$ is an infinitesimal vector displacement: you go an infinitesimal distance dr in whatever direction $d\mathbf{r}$ points. In Cartesian coordinates, for example, such a displacement would in general be in some combination of the x, y, and z directions: you would move from (x, y, z) to $(x+dx, y+dy, z+dz)$ by moving dx in the $\widehat{\mathbf{x}}$ direction, dy in the $\widehat{\mathbf{y}}$ direction, and dz in the $\widehat{\mathbf{z}}$ direction.[2] The corresponding line element would thus be

$$d\mathbf{r} = dx\,\widehat{\mathbf{x}} + dy\,\widehat{\mathbf{y}} + dz\,\widehat{\mathbf{z}} \tag{2.1}$$

Just as $\widehat{\mathbf{x}}$, $\widehat{\mathbf{y}}$, and $\widehat{\mathbf{z}}$ point in the positive x, positive y, and positive z directions, in two-dimensional polar coordinates (r, θ), $\widehat{\mathbf{r}}$ points in the positive r and $\widehat{\boldsymbol{\theta}}$ in the positive θ direction. That is, $\widehat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$ point in the direction of increasing r and θ, respectively, so that $\widehat{\mathbf{r}}$ points radially outward and $\widehat{\boldsymbol{\theta}}$ is in the counterclockwise tangential direction, as shown in fig. (2.2).[3] The line element $d\mathbf{r}$ thus has two contributions: one from the displacement dr along the radial direction and one from the displacement $r\,d\theta$ along the arc in the tangential direction:

$$d\mathbf{r} = dr\,\widehat{\mathbf{r}} + r\,d\theta\,\widehat{\boldsymbol{\theta}} \tag{2.2}$$

To extend this to three-dimensional cylindrical coordinates (r, θ, z), we need only add in the contribution from the displacement dz in the z direction:

$$d\mathbf{r} = dr\,\widehat{\mathbf{r}} + r\,d\theta\,\widehat{\boldsymbol{\theta}} + dz\,\widehat{\mathbf{z}} \tag{2.3}$$

To get to the point specified by (r, θ, φ) in spherical coordinates,[4] you first go a distance r up the z axis, then swing through the polar angle θ toward the x axis, then swing through azimuthal angle φ in the xy plane (in a right-handed sense about the z axis). As more or less shown in fig. (2.3),

[2] Any of dx, dy, or dz of course can and will be negative if you are moving in the negative x, y, or z direction.

[3] This prescription is of course very general: for any coordinate ξ, $\widehat{\boldsymbol{\xi}}$ is in the direction of increasing ξ, that is, in the direction that a positive $d\xi$ will take us.

[4] The notation that physicists use for the polar and azimuthal angles is generally the reverse of that used by math people. You just have to get used to it. The world is an imperfect place.

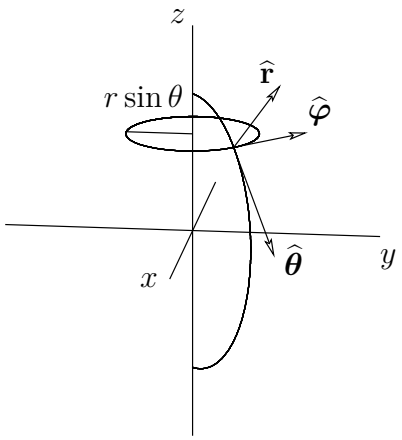


Figure 2.3: Arc Lengths in Spherical Coordinates

the arc length ds_θ corresponding to $d\theta$ involves the full radius r, while the arc length ds_φ corresponding to $d\varphi$ involves the radius $r\sin\theta$: $ds_\theta = r\,d\theta$, $ds_\varphi = r\sin\theta\,d\varphi$. The line element in spherical coordinates is therefore

$$d\mathbf{r} = dr\,\hat{\mathbf{r}} + r\,d\theta\,\hat{\boldsymbol{\theta}} + r\sin\theta\,d\varphi\,\hat{\boldsymbol{\varphi}} \tag{2.4}$$

The expressions (2.1), (2.2), (2.3), and (2.4) for the line element $d\mathbf{r}$ are useful for carrying out line integrations like

$$\int_P^{\mathbf{r}_2} \mathbf{F}\cdot d\mathbf{r}$$
$$\mathbf{r}_1$$

where P is the path followed by a body as it travels from location $\mathbf{r}_1$ to location $\mathbf{r}_2$ and $\mathbf{F}$ is some vector quantity. If, for example, $\mathbf{F}$ is the force exerted on the body, then the above line integral would, as we will see in Chapter 5, give the work W done by that force on the body.

Suppose, for example, that the force $\mathbf{F} = -mg\hat{\mathbf{y}}$, where m and g are constants. If we want the work done by this force as the body travels along a straight line from $(0, a)$ to $(a, 0)$, as shown in fig. (2.4), we have [5]

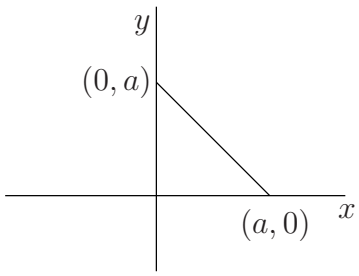


Figure 2.4: The Path Followed by the Body

$$\begin{aligned} W &= \int_{(0,a)}^{(a,0)} \mathbf{F} \cdot d\mathbf{r} \\ &= \int_{(0,a)}^{(a,0)} -mg\widehat{\mathbf{y}} \cdot (dx\,\widehat{\mathbf{x}} + dy\,\widehat{\mathbf{y}}) \\ &= \int_{(0,a)}^{(a,0)} (-mg\,dx\,\widehat{\mathbf{y}} \cdot \widehat{\mathbf{x}} - mg\,dy\,\widehat{\mathbf{y}} \cdot \widehat{\mathbf{y}}) \end{aligned}$$

Since $\widehat{\mathbf{y}} \cdot \widehat{\mathbf{x}} = 0$ and $\widehat{\mathbf{y}} \cdot \widehat{\mathbf{y}} = 1$, this reduces to

$$\begin{aligned} W &= \int_{(0,a)}^{(a,0)} -mg\,dy \\ &= -mg \int_{(0,a)}^{(a,0)} dy \\ &= -mg[y]_a^0 \\ &= mga \end{aligned}$$

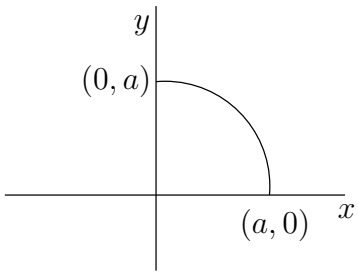


Figure 2.5: Another Path Followed by the Body

Now suppose that we want the work done by this same force as the body travels along the circular arc of fig. (2.5) between those same two points. We could carry out the integration exactly as we did for the first path, so that again $W = mga$. In fact, since $\mathbf{F}$ is constant [6] and can therefore be pulled outside of the integration, the line integral would work out to the same result for *any* path between the points $(0, a)$ and $(a, 0)$:

$$\begin{aligned} W &= \int_{(0,a)}^{(a,0)} \mathbf{F} \cdot d\mathbf{r} \\ &= \mathbf{F} \cdot \int_{(0,a)}^{(a,0)} d\mathbf{r} \end{aligned}$$

[5] We could have noted that along the path we are taking $dy = -dx$, but we can carry out this particular integration without making use of this fact.

[6] Note that since $\mathbf{F}$ is a vector, this means being constant both in magnitude and in direction.

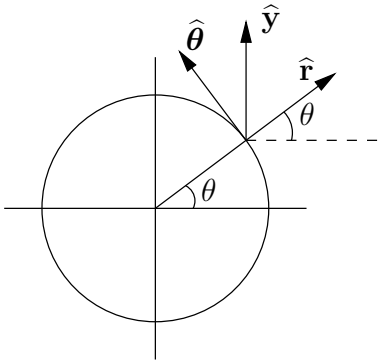


Figure 2.6: People Aren't Wearing Enough Hats

$$
\begin{aligned}
&= \mathbf{F} \cdot \Big[\mathbf{r}\Big]_{(0,a)}^{(a,0)} \\
&= \mathbf{F} \cdot \Big[x\,\widehat{\mathbf{x}} + y\,\widehat{\mathbf{y}}\Big]_{(0,a)}^{(a,0)} \\
&= \mathbf{F} \cdot \Big([x]_0^a\;\widehat{\mathbf{x}} + [y]_a^0\;\widehat{\mathbf{y}}\Big) \\
&= \mathbf{F} \cdot (a\,\widehat{\mathbf{x}} - a\,\widehat{\mathbf{y}}) \\
&= -mg\,\widehat{\mathbf{y}} \cdot (a\,\widehat{\mathbf{x}} - a\,\widehat{\mathbf{y}}) \\
&= mga
\end{aligned}
$$

For the sake of an exercise in polar coordinates, however, let us actually set up $d\mathbf{r}$ along the circular arc and do out the integration. Using the polar $d\mathbf{r}$ of eq.(2.2), we have (changing our notation for the endpoints of the integration to $\mathbf{r}_i$ and $\mathbf{r}_f$)

$$
\begin{aligned}
W &= \int_{\mathbf{r}_i}^{\mathbf{r}_f} \mathbf{F} \cdot d\mathbf{r} \\
&= \int_{\mathbf{r}_i}^{\mathbf{r}_f} -mg\,\widehat{\mathbf{y}} \cdot (dr\,\widehat{\mathbf{r}} + r\,d\theta\,\widehat{\boldsymbol{\theta}}) \\
&= -mg \int_{\mathbf{r}_i}^{\mathbf{r}_f} (dr\,\widehat{\mathbf{y}} \cdot \widehat{\mathbf{r}} + r\,d\theta\,\widehat{\mathbf{y}} \cdot \widehat{\boldsymbol{\theta}})
\end{aligned}
$$

As you can see from fig.(2.6), the angle between $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{r}}$ is $\frac{\pi}{2} - \theta$ and the angle between $\widehat{\mathbf{y}}$ and $\widehat{\boldsymbol{\theta}}$ is θ. Thus

$$
\begin{aligned}
\widehat{\mathbf{y}} \cdot \widehat{\mathbf{r}} &= 1(1)\cos(\frac{\pi}{2} - \theta) = \sin\theta \\
\widehat{\mathbf{y}} \cdot \widehat{\boldsymbol{\theta}} &= 1(1)\cos(\theta) = \cos\theta
\end{aligned}
$$

and we have

$$
W = -mg \int_{\mathbf{r}_i}^{\mathbf{r}_f} (dr\sin\theta + r\,d\theta\cos\theta)
$$

Along the circular arc the body is following, $r = a$, which, being a constant, means that $dr = 0$. So also $r\,d\theta = a\,d\theta$, with θ going from $\frac{\pi}{2}$ to 0. Thus

$$\begin{aligned} W &= -mg \int_{\mathbf{r}_i}^{\mathbf{r}_f} (0 + a\,d\theta \cos\theta) \\ &= -mg \int_{\frac{\pi}{2}}^{0} a\,d\theta \cos\theta \\ &= -mga \int_{\frac{\pi}{2}}^{0} \cos\theta\,d\theta \\ &= -mga \Big[\sin\theta \Big]_{\frac{\pi}{2}}^{0} \\ &= mga \end{aligned}$$

2.2 Surface Elements & Integrations

Surface elements are infinitesimal patches aligned with the coordinate planes. In Cartesian coordinates, for example, the surface elements are infinitesimal rectangles parallel to the xy, yz, and xz planes. Fig. (2.7) shows such a patch in the xy plane: the infinitesimal rectangle that extends from (x, y) to $(x + dx, y + dy)$. The area dA of this patch is simply $dx\,dy$. The areas of patches parallel to the yz and xz planes are similarly $dy\,dz$ and $dx\,dz$.

In curvilinear coordinate systems (polar, cylindrical, spherical, etc.), patches with sides along the coordinate axes are in general not rectangular. Fig. (2.8) shows, for example, the shape of patches in polar coordinates: both sides along the radial direction are straight and of length dr, but the other two sides are circular arcs, with the outer longer than the inner. Since these patches are infinitesimal, however, the deviations from rectangularity will be second-order small and may be neglected: such corrections would be infinitesimal even in comparison to the infinitesimal areas of the patches.[7] We

[7] For the benefit of those of you neurotically incapable of trusting others, the exact area

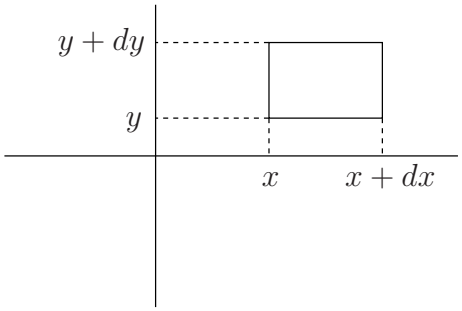


Figure 2.7: A Cartesian Surface Element

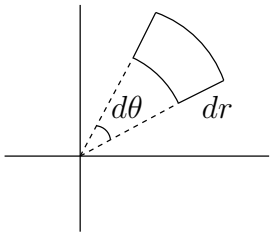


Figure 2.8: The Polar Area Element

may therefore treat the patch shown in fig. (2.8) as though it were a rectangle of length dr on one side and $r\,d\theta$ on the other, so that

$$dA = (dr)(r\,d\theta) = r\,dr\,d\theta$$

The general procedure for obtaining results for area elements in the various coordinate systems is simply to think of the sides of the patches in question — their arc lengths, as given in the relations for line elements in the preceding section — and then simply multiply length by width to get the area element:

In cylindrical coordinates (r, θ, z), patches in the $r\theta$ plane are the same as the polar patches we just did: $dA = r\,dr\,d\theta$. Patches on cylindrical surfaces around the z axis will be $r\,d\theta$ in the direction around the cylinder and dz in the direction along the axis of the cylinder: $dA = r\,d\theta\,dz$.

In spherical coordinates (r, θ, φ), patches on spherical surfaces centered at the origin will be $r\,d\theta$ on one side and $r\sin\theta\,d\varphi$ along the other: $dA = r^2\sin\theta\,d\theta\,d\varphi$.

Although $\sin\theta\,d\theta\,d\varphi$ is a perfectly fine expression in both the æsthetic and utilitarian senses, just so that you are aware of it, this combination is also often expressed as $d\Omega$, where Ω is the solid angle. Just as we see from integrating the line element $ds = r\,d\theta$ that there is an angle of 2π in a full circle ($s = \theta r$ with s being the full circumference $2\pi r$), we see from integrating $dA = r^2 d\Omega$ that there is a total solid angle of 4π in a sphere ($A = r^2\Omega$ with A being the full area $4\pi r^2$). More explicitly, in terms of the angles θ and φ,

$$A = \int_{\theta=0,\ \varphi=0}^{\theta=\pi,\ \varphi=2\pi} r^2\sin\theta\,d\theta\,d\varphi$$

of the patch shown in fig. (2.8) would be the fraction $d\theta/2\pi$ times the area of an annulus (washer) of inner radius r and outer radius $r + dr$, so that

$$dA = \big(\pi(r+dr)^2 - \pi r^2\big)\frac{d\theta}{2\pi} = r\,dr\,d\theta + \tfrac{1}{2}dr^2 d\theta \approx r\,dr\,d\theta$$

$$\begin{aligned} &= r^2 \int_0^{\pi} \sin\theta \, d\theta \int_0^{2\pi} d\varphi \\ &= r^2 \Big[-\cos\theta \Big]_0^{\pi} (2\pi) \\ &= 4\pi r^2 \end{aligned}$$

As a practical example of a surface integration, we will work out the moment of inertia I of a uniform circular disk of radius a and mass m, which turns out to be given by

$$I = \int_{\text{disk}} r^2 \frac{m}{\pi a^2} \, dA$$

Using $dA = r \, dr \, d\theta$ and noting that as we integrate over the disk, r goes from 0 to a and θ from 0 to 2π, we have

$$\begin{aligned} I &= \int_{\text{disk}} r^2 \frac{m}{\pi a^2} \, r \, dr \, d\theta \\ &= \frac{m}{\pi a^2} \int_0^a r^3 \, dr \int_0^{2\pi} d\theta \\ &= \frac{m}{\pi a^2} (\tfrac{1}{4} a^4)(2\pi) \\ &= \tfrac{1}{2} m a^2 \end{aligned}$$

2.3 Volume Elements & Integrations

Just as it was possible to treat surface elements, even in curvilinear coordinates, as rectangles, it is possible to treat volume elements as rectangular boxes, with the lengths of the sides being the arc lengths given by the corresponding components of the line elements. The volume element in Cartesian coordinates is thus simply the product of dx, dy, and dz:

$$dV = dx \, dy \, dz$$

In cylindrical coordinates, the boxes will be dr along the radial side, $r \, d\theta$ along the $\widehat{\boldsymbol{\theta}}$ side, and dz along the side parallel to the z axis:

$$dV = (dr)(r \, d\theta)(dz) = r \, dr \, d\theta \, dz$$

And in spherical coordinates, the sides are dr, $r \, d\theta$, and $r \sin\theta \, d\varphi$:

$$dV = (dr)(r \, d\theta)(r \sin\theta \, d\varphi) = r^2 \sin\theta \, dr \, d\theta \, d\varphi$$

Suppose, for example, that we are trying to find the location of the center of mass of a uniform hemisphere of radius a and mass m that is centered at the

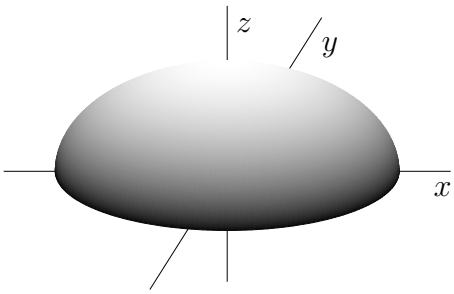


Figure 2.9: A Sort of Hemisphere

origin and with its base in the xy plane and facing the negative z direction, as shown, rather pathetically, but at the absolute outer limit of our graphical abilities, in fig. (2.9). The relation for the z coordinate z_{cm} of its center of mass turns out to be given by

$$z_{cm} = \frac{1}{m} \int\limits_{\text{hemisphere}} z \, \frac{m}{\frac{2}{3}\pi a^3} \, dV$$

Using $dV = r^2 \sin\theta \, dr \, d\theta \, d\varphi$ and $z = r\cos\theta$, and noting that, as we integrate over the hemisphere, r goes from 0 to a, θ from 0 to $\frac{\pi}{2}$, and φ from 0 to 2π, we have

$$\begin{aligned}
z_{cm} &= \frac{1}{m} \int\limits_{\text{hemisphere}} z \, \frac{m}{\frac{2}{3}\pi a^3} \, r^2 \sin\theta \, dr \, d\theta \, d\varphi \\
&= \frac{1}{m} \, \frac{m}{\frac{2}{3}\pi a^3} \int\limits_{\text{hemisphere}} r\cos\theta \; r^2 \sin\theta \, dr \, d\theta \, d\varphi \\
&= \frac{1}{m} \, \frac{m}{\frac{2}{3}\pi a^3} \int_0^a r^3 \, dr \int_0^{\frac{\pi}{2}} \cos\theta \sin\theta \, d\theta \int_0^{2\pi} d\varphi \\
&= \frac{1}{m} \, \frac{m}{\frac{2}{3}\pi a^3} \, \tfrac{1}{4}a^4 \int_0^{\frac{\pi}{2}} \tfrac{1}{2}\sin(2\theta) \, d\theta \; 2\pi \\
&= \frac{1}{m} \, \frac{m}{\frac{2}{3}\pi a^3} \, \tfrac{1}{4}a^4 \left[-\tfrac{1}{4}\cos(2\theta) \right]_0^{\frac{\pi}{2}} \; 2\pi \\
&= \frac{1}{m} \, \frac{m}{\frac{2}{3}\pi a^3} \, \tfrac{1}{4}a^4 \, \tfrac{1}{2} \, 2\pi \\
&= \tfrac{3}{8}a
\end{aligned}$$

2.4 The Gradient

The gradient operator is denoted by a ∇ ("del") and, in Cartesian coordinates, is given by

$$\nabla = \frac{\partial}{\partial x}\,\hat{\mathbf{x}} + \frac{\partial}{\partial y}\,\hat{\mathbf{y}} + \frac{\partial}{\partial z}\,\hat{\mathbf{z}}$$

Why the gradient operator should be defined this way, and what its uses and significance are, will remain a profound mystery for the moment. At this point it will suffice to note that when it is applied to a scalar function $U = U(x, y, z)$, it gives for the gradient ∇U of U

$$\nabla U = \frac{\partial U}{\partial x}\,\hat{\mathbf{x}} + \frac{\partial U}{\partial y}\,\hat{\mathbf{y}} + \frac{\partial U}{\partial z}\,\hat{\mathbf{z}}$$

That is, ∇U is a vector, the components of which are

$$(\nabla U)_x = \frac{\partial U}{\partial x} \qquad (\nabla U)_y = \frac{\partial U}{\partial y} \qquad (\nabla U)_z = \frac{\partial U}{\partial z}$$

It is also possible to express the gradient operator in curvilinear coordinate systems (cylindrical, spherical, etc.), but we don't need to get into that.

For those of you who've never dealt with partial derivatives like $\partial/\partial x$ before and are wondering what the difference is between the partial derivative $\partial/\partial x$ and the ordinary derivative (technically, the total derivative) d/dx, the answer is: not much. For functions of one variable, like $f(x)$, taking the derivative means taking df/dx — there's only one variable with respect to which you can take a derivative. With functions of several variables, like $U(x, y, z)$, however, there are three variables with respect to which you can take a derivative — x, y, or z. Taking the partial derivative of U with respect to x means simply taking the derivative with respect to x *treating* y *and* z *as constants*. So if $U = 4\pi(x^2 + y)z$,

$$\frac{\partial U}{\partial x} = 4\pi(2x)z = 8\pi xz$$

One important relation involving the gradient operator is

$$dU = \nabla U \cdot d\mathbf{r} \tag{2.5}$$

That is, the change dU in the function U as we take an infinitesimal step $d\mathbf{r}$ from location $\mathbf{r}$ to location $\mathbf{r} + d\mathbf{r}$ is given by $\nabla U \cdot d\mathbf{r}$. To establish this relation, we simply expand the right-hand side in Cartesian coordinates:

$$dU \stackrel{?}{=} \nabla U \cdot d\mathbf{r} = \left(\frac{\partial U}{\partial x}\,\hat{\mathbf{x}} + \frac{\partial U}{\partial y}\,\hat{\mathbf{y}} + \frac{\partial U}{\partial z}\,\hat{\mathbf{z}}\right) \cdot (dx\,\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}} + dz\,\hat{\mathbf{z}})$$

All of the cross terms will involve vanishing dot products like $\hat{\mathbf{x}} \cdot \hat{\mathbf{y}} = 0$; only the direct terms, which involve dot products like $\hat{\mathbf{x}} \cdot \hat{\mathbf{x}} = 1$, will survive. Thus

$$dU \stackrel{?}{=} \frac{\partial U}{\partial x}\,dx + \frac{\partial U}{\partial y}\,dy + \frac{\partial U}{\partial z}\,dz$$

which is in fact true: it is simply stating that the total change dU in U is the sum of the changes in U due to the variations in each of the variables x, y, and z.

A noteworthy consequence of relation (2.5) is that the gradient of a function *points in the direction of the steepest rate of increase* of that function: by the geometric definition of the dot product,

$$dU = \nabla U \cdot d\mathbf{r} = |\nabla U|\,|d\mathbf{r}| \cos\theta$$

where θ is the angle between ∇U and $d\mathbf{r}$. This dot product, and therefore the change dU in U, will be greatest when $\theta = 0$ — that is, when ∇U and the displacement $d\mathbf{r}$ are in the same direction. This property of the gradient is in fact why it is called a gradient: if the function U is regarded as the "terrain," ∇U points mostly steeply "uphill."

A corollary of this is that a function does not change in value if you move in the direction perpendicular to its gradient: when $d\mathbf{r} \perp \nabla U$,

$$dU = |\nabla U|\,|d\mathbf{r}| \cos\theta = |\nabla U|\,|d\mathbf{r}| \cos\frac{\pi}{2} = 0$$

Since there is no change dU, U remains constant. So by always moving at right angles to the gradient of a function, we can trace out the level lines and level surfaces of the function, that is, the lines and surfaces along which the function is constant.

Consider, for example, the function $U = x^2 + y^2$. Since $r^2 = x^2 + y^2$, the value of U is just the square of our distance from the origin, so that a three-dimensional plot of U would look like a parabolic bowl with its bottom at the origin. The gradient of this function is

$$\begin{aligned}
\nabla U &= \frac{\partial}{\partial x}(x^2 + y^2)\,\hat{\mathbf{x}} + \frac{\partial}{\partial y}(x^2 + y^2)\,\hat{\mathbf{y}} \\
&= 2x\,\hat{\mathbf{x}} + 2y\,\hat{\mathbf{y}} \\
&= 2(x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}}) \\
&= 2\mathbf{r}
\end{aligned}$$

where we have noted that $x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}}$ is just the position vector $\mathbf{r}$. ∇U is thus in the same direction as $\mathbf{r}$ — radially away from the origin, which is indeed the steepest direction we can go along the side of the bowl. The direction perpendicular to $\mathbf{r}$, and thus to the gradient, is the tangential direction $\hat{\boldsymbol{\theta}}$, which means that the level lines of U are circles concentric with the origin. This is also just what we would have expected: since $U = x^2 + y^2 = r^2$, the lines of constant U are those along which the radius r is constant.

2.5 Divergence & Gauss's Theorem

The divergence of a vector function

$$\mathbf{E}(x,y,z) = E_x(x,y,z)\,\hat{\mathbf{x}} + E_y(x,y,z)\,\hat{\mathbf{y}} + E_z(x,y,z)\,\hat{\mathbf{z}}$$

is defined as $\nabla \cdot \mathbf{E}$. Thus

$$\begin{aligned}\nabla \cdot \mathbf{E} &= \left(\frac{\partial}{\partial x}\,\hat{\mathbf{x}} + \frac{\partial}{\partial y}\,\hat{\mathbf{y}} + \frac{\partial}{\partial z}\,\hat{\mathbf{z}}\right) \\ &\qquad \cdot \Big(E_x(x,y,z)\,\hat{\mathbf{x}} + E_y(x,y,z)\,\hat{\mathbf{y}} + E_z(x,y,z)\,\hat{\mathbf{z}}\Big) \\ &= \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}\end{aligned}$$

A very important theorem involving the divergence was proved by Gauss and is therefore known, aptly enough, as Gauss's theorem: [8]

$$\int_V dV\,\nabla \cdot \mathbf{E} = \oint_S dA\,\hat{\mathbf{n}} \cdot \mathbf{E} \tag{2.6}$$

On the left-hand side, the divergence of the vector field $\mathbf{E}$ is being integrated over a finite volume V. The right-hand side is an integral of the normal component of $\mathbf{E}$ over the surface S that encloses the volume V, with $\hat{\mathbf{n}}$ taken

[8] Here we begin writing integrals the big-people way, with the differential next to the integral sign and the integrand following it. Thus we will write, for example,

$$\int dx\, e^x \sin x$$

rather than

$$\int e^x \sin x\, dx$$

It is more sensible to write integrals this way for a couple of reasons: When the integration is multiple (that is, by more than one variable), it makes it unambiguous which limits go with which variable. (As opposed to the confusion created by

$$\int_0^a \int_0^b xy\, dx\, dy$$

where it is impossible to tell whether x is supposed to go from 0 to a and y from 0 to b or the other way around — it's like trying to figure out which fork to use at a formal dinner.

$$\int_0^a dx \int_0^b dy\; xy$$

is, in contrast, totally clear.) Also, in the real world integrands can be very complex — sometimes pages long. By writing the differential next to the integral sign, you can see right away which variable the integration is with respect to, without having to dig. So get used to writing your integrals this way.

by convention to be the outward normal to the surface S, that is, the unit vector that at each location on the surface points perpendicularly outward from it. The loop on the integral sign on the right-hand side is a conventional notation indicating that the integration is over a closed surface.

As a simple if rather artificial example, consider the case $\mathbf{E} = x^2\hat{\mathbf{x}}$ with V being a cube with one corner at the origin and the far corner at (a, a, a). The left-hand side of Gauss's theorem will work out to

$$\begin{aligned}
\int_V dV\, \nabla \cdot \mathbf{E} &= \int_{\text{cube}} dx\, dy\, dz \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \right) \\
&= \int_0^a dx \int_0^a dy \int_0^a dz \left(\frac{\partial (x^2)}{\partial x} + \frac{\partial (0)}{\partial y} + \frac{\partial (0)}{\partial z} \right) \\
&= \int_0^a dx \int_0^a dy \int_0^a dz\, 2x \\
&= \Big[x^2\Big]_0^a \Big[y\Big]_0^a \Big[z\Big]_0^a \\
&= a^2(a)(a) \\
&= a^4
\end{aligned}$$

The integration over the surface of the cube on the right-hand side of Gauss's theorem we will break into six separate integrations, one over each face of the cube. As shown in fig. (2.10), for the two faces parallel to the xy plane, the outward normal is $-\hat{\mathbf{z}}$ on the face at $z = 0$ and $+\hat{\mathbf{z}}$ on the face at $z = a$. Since $E_z = 0$, $\hat{\mathbf{n}} \cdot \mathbf{E} = 0$ on these faces. For the two faces parallel to the xz plane, the outward normal is $-\hat{\mathbf{y}}$ on the face at $y = 0$ and $+\hat{\mathbf{y}}$ on the face at $y = a$. Since $E_y = 0$, $\hat{\mathbf{n}} \cdot \mathbf{E} = 0$ on these faces. For the two faces parallel to the yz plane, the outward normal is $-\hat{\mathbf{x}}$ on the face at $x = 0$ and $+\hat{\mathbf{x}}$ on the

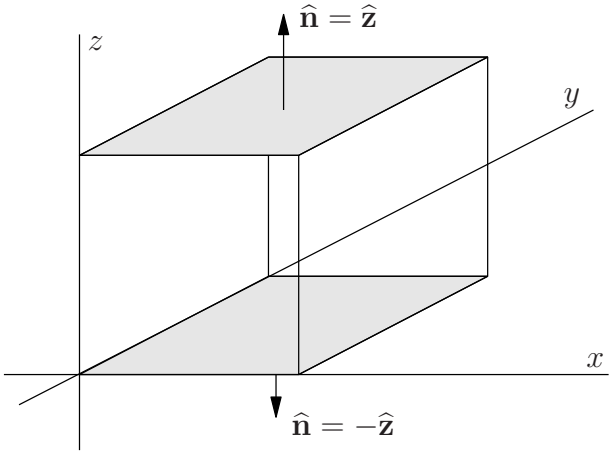


Figure 2.10: A Cube!

face at $x = a$. There is no contribution from the face at $x = 0$, since on this face $E_x = x^2 = 0$. The only nonzero contribution to the right-hand side of Gauss's theorem comes from the face at $x = a$, on which $E_x = x^2 = a^2$. The contribution of this face is

$$\begin{aligned}\int\limits_{\text{face at } x=a} dA\,\hat{\mathbf{n}}\cdot\mathbf{E} &= \int\limits_{\text{face at } x=a} dy\,dz\,\hat{\mathbf{x}}\cdot(a^2\hat{\mathbf{x}}) \\ &= \int_0^a dy \int_0^a dz\,a^2 \\ &= \Big[y\Big]_0^a \Big[z\Big]_0^a a^2 \\ &= a(a)(a^2) \\ &= a^4\end{aligned}$$

So the right-hand and left-hand sides of Gauss's theorem are indeed equal for this particular example.

The $\hat{\mathbf{n}}\cdot\mathbf{E}$ that occurs on the right-hand side of Gauss's theorem is the component of $\mathbf{E}$ pointing perpendicularly outward from the surface. When integrated over the surface S, it gives a measure of how much of $\mathbf{E}$ is passing through the surface and is known as the flux of $\mathbf{E}$ through S. On the left-hand side of Gauss's theorem, this is related to the integral of the divergence of $\mathbf{E}$ throughout the enclosed volume. If applied to an infinitesimal volume V and enclosing surface S, Gauss's theorem is saying that the value of $\nabla\cdot\mathbf{E}$ at a point is a measure of the flux of $\mathbf{E}$ out of that point. This is in fact why $\nabla\cdot\mathbf{E}$ is called the divergence.

To prove Gauss's theorem, we consider an infinitesimal rectangular box aligned with the coordinate axes, with far corners at (x, y, z) and $(x+dx, y+dy, z+dz)$. Since the box is infinitesimal, there is only a single contribution to the left-hand side of Gauss's theorem:

$$dV\,\nabla\cdot\mathbf{E} = dx\,dy\,dz\left(\frac{\partial E_x}{\partial x}+\frac{\partial E_y}{\partial y}+\frac{\partial E_z}{\partial z}\right) \tag{2.7}$$

On the right-hand side of Gauss's theorem, there are six infinitesimal contributions, one from each face of the box, just as in the above example. If we denote these faces as the x face, the $x+dx$ face, etc., the right-hand side of Gauss's theorem expands to

$$\begin{aligned}\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} = &\Big[dA\,\hat{\mathbf{n}}\cdot\mathbf{E}\Big]_{\text{face at } x} + \Big[dA\,\hat{\mathbf{n}}\cdot\mathbf{E}\Big]_{\text{face at } x+dx} \\ &+ \Big[dA\,\hat{\mathbf{n}}\cdot\mathbf{E}\Big]_{\text{face at } y} + \Big[dA\,\hat{\mathbf{n}}\cdot\mathbf{E}\Big]_{\text{face at } y+dy} \\ &+ \Big[dA\,\hat{\mathbf{n}}\cdot\mathbf{E}\Big]_{\text{face at } z} + \Big[dA\,\hat{\mathbf{n}}\cdot\mathbf{E}\Big]_{\text{face at } z+dz}\end{aligned}$$

Using the same results for the normal $\hat{\mathbf{n}}$ that we did in the example, this becomes

$$\begin{aligned}
\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} &= \Big[dy\,dz\,(-\hat{\mathbf{x}})\cdot\mathbf{E}\Big]_{\text{face at } x} \;+\; \Big[dy\,dz\,\hat{\mathbf{x}}\cdot\mathbf{E}\Big]_{\text{face at } x+dx} \\
&\quad+\; \Big[dx\,dz\,(-\hat{\mathbf{y}})\cdot\mathbf{E}\Big]_{\text{face at } y} \;+\; \Big[dx\,dz\,\hat{\mathbf{y}}\cdot\mathbf{E}\Big]_{\text{face at } y+dy} \\
&\quad+\; \Big[dx\,dy\,(-\hat{\mathbf{z}})\cdot\mathbf{E}\Big]_{\text{face at } z} \;+\; \Big[dx\,dy\,\hat{\mathbf{z}}\cdot\mathbf{E}\Big]_{\text{face at } z+dz} \\
&= \Big[dy\,dz\,(-E_x)\Big]_{\text{face at } x} \;+\; \Big[dy\,dz\,E_x\Big]_{\text{face at } x+dx} \\
&\quad+\; \Big[dx\,dz\,(-E_y)\Big]_{\text{face at } y} \;+\; \Big[dx\,dz\,E_y\Big]_{\text{face at } y+dy} \\
&\quad+\; \Big[dx\,dy\,(-E_z)\Big]_{\text{face at } z} \;+\; \Big[dx\,dy\,E_z\Big]_{\text{face at } z+dz} \\
&= -dy\,dz\,E_x(x) + dy\,dz\,E_x(x+dx) \\
&\quad- dx\,dz\,E_y(y) + dx\,dz\,E_y(y+dy) \\
&\quad- dx\,dy\,E_z(z) + dx\,dy\,E_z(z+dz) \\
&= \Big(E_x(x+dx) - E_x(x)\Big)dy\,dz \\
&\quad+ \Big(E_y(y+dy) - E_y(y)\Big)dx\,dz \\
&\quad+ \Big(E_z(z+dz) - E_z(z)\Big)dy\,dz
\end{aligned}$$

Now, $E_x(x+dx) - E_x(x)$ is just the change in E_x as we go from x to $x+dx$ — a change we could equally well write as

$$\frac{\partial E_x}{\partial x}\,dx$$

The right-hand side of Gauss's theorem thus reduces to

$$\begin{aligned}
\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} &= \left(\frac{\partial E_x}{\partial x}\,dx\right)dy\,dz + \left(\frac{\partial E_y}{\partial y}\,dy\right)dx\,dz + \left(\frac{\partial E_z}{\partial z}\,dz\right)dy\,dz \\
&= \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z}\right)dx\,dy\,dz
\end{aligned}$$

which is the same as what the left-hand side worked out to.

We have now established the validity of Gauss's theorem for infinitesimal rectangular boxes aligned with the coordinate axes. To complete the proof of the theorem, we note that any finite volume and its enclosing surface can be built up out of such infinitesimal boxes. The volume integral on the left-hand side of Gauss's theorem will then be simply the sum of the contributions from the volumes of the infinitesimal boxes. In the surface integral on the right-hand side of Gauss's theorem, however, we will have the

sum over all the faces of all of the boxes, which, since many of these faces are interior to the volume, would seem at first not to be what we want. In the interior of the volume, however, the boxes fit flush against each other, so that their faces abut and they occur in pairs with their outward normals pointing in opposite directions. Contributions from these interior faces will therefore cancel each other out when summed over. The only faces that will survive this cancellation are those on the exterior, which together make up the enclosing surface S. Word.

2.6 Curl & Stokes's Theorem

The curl of a vector function

$$\mathbf{E}(x,y,z) = E_x(x,y,z)\,\hat{\mathbf{x}} + E_y(x,y,z)\,\hat{\mathbf{y}} + E_z(x,y,z)\,\hat{\mathbf{z}}$$

is defined as $\nabla \times \mathbf{E}$. Thus

$$\begin{aligned}
\nabla \times \mathbf{E} &= \left(\frac{\partial}{\partial x}\,\hat{\mathbf{x}} + \frac{\partial}{\partial y}\,\hat{\mathbf{y}} + \frac{\partial}{\partial z}\,\hat{\mathbf{z}}\right) \\
&\qquad \times \left(E_x(x,y,z)\,\hat{\mathbf{x}} + E_y(x,y,z)\,\hat{\mathbf{y}} + E_z(x,y,z)\,\hat{\mathbf{z}}\right) \\
&= \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ E_x & E_y & E_z \end{vmatrix} \\
&= \left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z}\right)\hat{\mathbf{x}} - \left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z}\right)\hat{\mathbf{y}} + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right)\hat{\mathbf{z}}
\end{aligned}$$

A very important theorem involving the curl was proved by Stokes and is known, as you would probably guess, as Stokes's theorem:

$$\oint_C \mathbf{E}\cdot d\mathbf{r} = \int_S dA\,\hat{\mathbf{n}}\cdot\nabla\times\mathbf{E} \tag{2.8}$$

In the line integral on the left-hand side of Stokes's theorem, the tangential component of the vector field $\mathbf{E}$ (that is, the component of $\mathbf{E}$ parallel to each infinitesimal displacement $d\mathbf{r}$) is being integrated around a closed curve C. The loop on the integral sign on the left-hand side is a conventional notation indicating that the integration is over a closed loop. The right-hand side is an integral of the normal component of $\nabla\times\mathbf{E}$ over *any* surface S that spans the loop C, with $\hat{\mathbf{n}}$ being the normal to the surface S. In Gauss's theorem, $\hat{\mathbf{n}}$ was taken to be the outward normal, but in Stokes's theorem the surface is not closed, so the surface has no inside or outside and we need some other way of specifying which of the two possible directions we should choose for the normal to the surface. The convention is to take $\hat{\mathbf{n}}$ to be in a right-handed

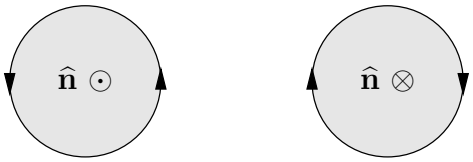


Figure 2.11: Directions in Stokes's Theorem

sense relative to the direction in which we are integrating around the loop on the left-hand side of Stokes's theorem. This convention is illustrated in fig. (2.11) for the case of a circular loop in the plane of the page and the circular disk that directly spans that loop: if we integrate around the loop counterclockwise, the normal to the disk should be out of the page; if we integrate around the loop clockwise, the normal should be into the page.

As a simple if rather artificial example, consider the case $\mathbf{E} = x\hat{\mathbf{y}}$ with C being a square with one corner at the origin and the far corner at $(a, a, 0)$ and S being the square surface that directly spans this square loop. If we go around the square counterclockwise,[9] as shown in fig. (2.12), the normal to the loop should be out of the page (in the $+\hat{\mathbf{z}}$ direction), so that the right-hand side of Stokes's theorem will work out to

$$\begin{aligned}
\int_S dA\, \hat{\mathbf{n}} \cdot \nabla \times \mathbf{E} &= \int_S dx\, dy\, \hat{\mathbf{z}} \cdot \left[\left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \right) \hat{\mathbf{x}} \right. \\
&\qquad \left. - \left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z} \right) \hat{\mathbf{y}} + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \right) \hat{\mathbf{z}} \right] \\
&= \int_S dx\, dy \left[\left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \right) \hat{\mathbf{z}} \cdot \hat{\mathbf{x}} \right. \\
&\qquad \left. - \left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z} \right) \hat{\mathbf{z}} \cdot \hat{\mathbf{y}} + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \right) \hat{\mathbf{z}} \cdot \hat{\mathbf{z}} \right] \\
&= \int_S dx\, dy \left[\left(\frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \right) (0) \right. \\
&\qquad \left. - \left(\frac{\partial E_z}{\partial x} - \frac{\partial E_x}{\partial z} \right) (0) + \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \right) (1) \right] \\
&= \int_S dx\, dy \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} \right) \\
&= \int_S dx\, dy \left(\frac{\partial (x)}{\partial x} - \frac{\partial (0)}{\partial y} \right)
\end{aligned}$$

[9] Clockwise and counterclockwise of course depend on the side from which you are looking. We mean counterclockwise when you are looking from "above" in the sense of looking "down" on the xy plane from the positive z axis.

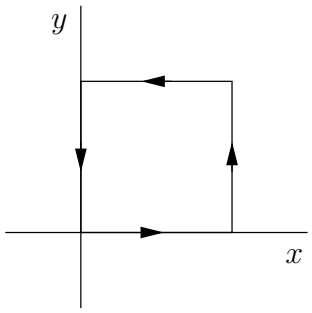


Figure 2.12: Going Around the Square Counterclockwise

$$= \int_0^a dx \int_0^a dy\,(1)$$
$$= a^2$$

The line integral around the perimeter of the square on the left-hand side of Stokes's theorem we will break into four separate integrations, one over each side of the square. For the two sides parallel to the x axis, $d\mathbf{r}$ is $+dx\,\hat{\mathbf{x}}$ on the side at $y = 0$ and $-dx\,\hat{\mathbf{x}}$ on the side at $y = a$, as you can see from fig. (2.12). Since $E_x = 0$, $\mathbf{E} \cdot d\mathbf{r} = 0$ on these sides. For the two sides parallel to the y axis, $d\mathbf{r}$ is $-dy\,\hat{\mathbf{y}}$ on the side at $x = 0$ and $+dy\,\hat{\mathbf{y}}$ on the side at $x = a$. There is no contribution from the side at $x = 0$, since on this side $E_y = x = 0$. The only nonzero contribution to the left-hand side of Stokes's theorem comes from the side at $x = a$, on which $E_y = x = a$. The contribution of this side is

$$\int_{\substack{\text{side at}\\ x=a}} \mathbf{E} \cdot d\mathbf{r} = \int_{\substack{\text{side at}\\ x=a}} a\,dy = \int_0^a dy\,a = \Big[y\Big]_0^a a = a(a) = a^2$$

So the right-hand and left-hand sides of Stokes's theorem are indeed equal for this particular example.

The line integral on the left-hand side of Stokes's theorem gives a measure of the extent to which the vector field $\mathbf{E}$ circulates around the loop C: the more vortex-like $\mathbf{E}$, the greater its tangential component and the greater the result of the line integral. This circulation or vorticity of $\mathbf{E}$ is related, on the right-hand side of Stokes's theorem, to the curl of $\mathbf{E}$ (at least to its normal component over the spanning surface). It is for this reason that $\nabla \times \mathbf{E}$ is called the curl. At any given point, $\hat{\mathbf{n}} \cdot \nabla \times \mathbf{E}$ will be greatest when $\hat{\mathbf{n}}$ is in the same direction as $\nabla \times \mathbf{E}$, so that the vorticity of $\mathbf{E}$ is greatest in the plane perpendicular to $\nabla \times \mathbf{E}$. In other words, the direction of $\nabla \times \mathbf{E}$ indicates, in a right-handed sense, the plane in which the vorticity of $\mathbf{E}$ is greatest.

To prove Stokes's theorem, we consider an infinitesimal rectangle aligned with the x and y axes, with far corners at (x, y, z) and $(x + dx, y + dy, z)$. Since the rectangle is infinitesimal, there is only a single contribution to the

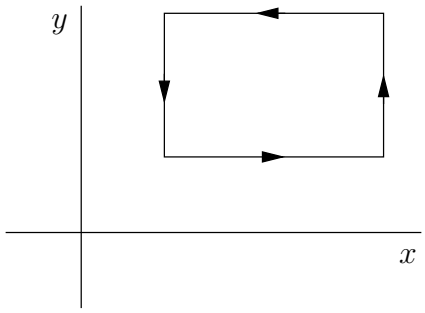


Figure 2.13: Directions for $d\mathbf{r}$

right-hand side of Stokes's theorem. If we go around the rectangle counterclockwise,[10] then the normal $\hat{\mathbf{n}}$ should be in the $+\hat{\mathbf{z}}$ direction and the right-hand side of Stokes's theorem reduces to

$$\begin{aligned} dA\,\hat{\mathbf{n}}\cdot\nabla\times\mathbf{E} &= dx\,dy\,\hat{\mathbf{z}}\cdot\left[\left(\frac{\partial E_z}{\partial y}-\frac{\partial E_y}{\partial z}\right)\hat{\mathbf{x}}\right. \\ &\qquad \left.-\left(\frac{\partial E_z}{\partial x}-\frac{\partial E_x}{\partial z}\right)\hat{\mathbf{y}}+\left(\frac{\partial E_y}{\partial x}-\frac{\partial E_x}{\partial y}\right)\hat{\mathbf{z}}\right] \\ &= dx\,dy\left(\frac{\partial E_y}{\partial x}-\frac{\partial E_x}{\partial y}\right) \end{aligned}$$

On the left-hand side of Stokes's theorem, there are four infinitesimal contributions, one from each side of the rectangle, just as in the above example. If we denote these sides as the x side, the $x+dx$ side, the y side, and the $y+dy$ side, and if we use the directions for $d\mathbf{r}$ indicated in fig. (2.13), the left-hand side of Stokes's theorem expands to

$$\begin{aligned} \oint_C \mathbf{E}\cdot d\mathbf{r} &= \Big[\mathbf{E}\cdot d\mathbf{r}\Big]_{\text{side at } x} + \Big[\mathbf{E}\cdot d\mathbf{r}\Big]_{\text{side at } x+dx} \\ &\qquad + \Big[\mathbf{E}\cdot d\mathbf{r}\Big]_{\text{side at } y} + \Big[\mathbf{E}\cdot d\mathbf{r}\Big]_{\text{side at } y+dy} \\ &= \Big[\mathbf{E}\cdot(-dy\,\hat{\mathbf{y}})\Big]_{\text{side at } x} + \Big[\mathbf{E}\cdot dy\,\hat{\mathbf{y}}\Big]_{\text{side at } x+dx} \\ &\qquad + \Big[\mathbf{E}\cdot dx\,\hat{\mathbf{x}}\Big]_{\text{side at } y} + \Big[\mathbf{E}\cdot(-dx\,\hat{\mathbf{x}})\Big]_{\text{side at } y+dy} \\ &= \Big[-E_y\,dy\Big]_{\text{side at } x} + \Big[E_y\,dy\Big]_{\text{side at } x+dx} \\ &\qquad + \Big[E_x\,dx\Big]_{\text{side at } y} + \Big[-E_x\,dx\Big]_{\text{side at } y+dy} \\ &= -E_y(x)\,dy + E_y(x+dx)\,dy \\ &\qquad + E_x(y)\,dx - E_x(y+dy)\,dx \end{aligned}$$

[10] With counterclockwise being defined the same way that it was in the example.

$$= \Big(E_y(x+dx) - E_y(x)\Big)\,dy - \Big(E_x(y+dy) - E_x(y)\Big)\,dx$$

Now, $E_y(x+dx) - E_y(x)$ is just the change in E_y as we go from x to $x+dx$ — a change we could equally well write as

$$\frac{\partial E_y}{\partial x}\,dx$$

Likewise

$$E_x(y+dy) - E_x(y) = \frac{\partial E_x}{\partial y}\,dy$$

The left-hand side of Stokes's theorem thus reduces to

$$\begin{aligned}\oint_C \mathbf{E}\cdot d\mathbf{r} &= \left(\frac{\partial E_y}{\partial x}\,dx\right)\,dy - \left(\frac{\partial E_x}{\partial y}\,dy\right)\,dx \\ &= \left(\frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}\right)dx\,dy\end{aligned}$$

which is the same as what the right-hand side worked out to.

We have now established Stokes's theorem for infinitesimal rectangles aligned with the x and y axes. To complete the proof of the theorem, we first note that we could equally well apply the above proof to rectangles aligned with the y and z axes or with the x and z axes. Then we note that any finite loop and its spanning surface can be built up out of such infinitesimal rectangles placed edge to edge. The surface integral on the right-hand side of Stokes's theorem will then be simply the sum of the contributions from the areas of the infinitesimal rectangles. In the line integral on the left-hand side of Stokes's theorem, however, we will have the sum over all the sides of all of the rectangles, which, since many of these sides are "interior" to the loop, would seem at first not to be what we want. In this interior region, however, the rectangles fit flush against each other, so that their sides abut and they occur in pairs with their line elements $d\mathbf{r}$ pointing in opposite directions. Contributions from these interior sides will therefore cancel each other out when summed over. The only sides that will survive this cancellation are those on the exterior, which together make up the loop C. Word.

Let us pause at this juncture for some silly pictures. Recall that the divergence $\nabla\cdot\mathbf{E}$ gives a measure of the extent to which a vector function $\mathbf{E}$ is diverging from (or, when $\nabla\cdot\mathbf{E} < 0$, converging into) a point, while $\nabla\times\mathbf{E}$ gives a measure of the extent to which that vector function is swirling or circulating about a point. Thus vector functions like those shown in fig. (2.14) are pure divergences, in the sense that they have nonzero divergences but zero curl, with the sign on that divergence being positive for functions "coming out of" a point and negative for functions "going into" a point. A vector

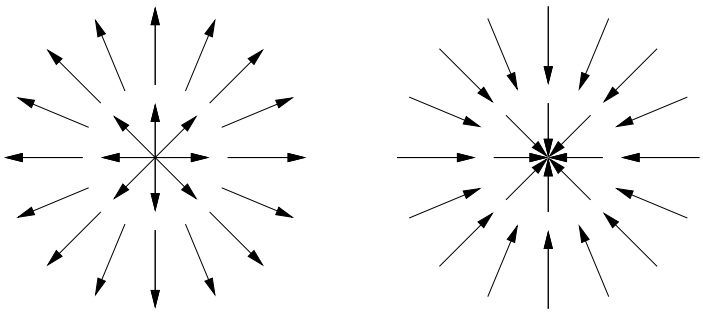


Figure 2.14: The Gist of a Pure Divergence

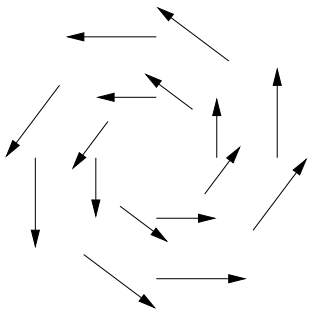


Figure 2.15: The Gist of a Pure Curl

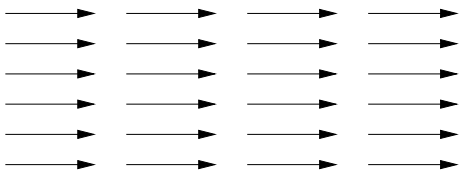


Figure 2.16: The Gist of Pure Nothing

function like that shown in fig. (2.15), on the other hand, is a pure curl, in the sense that it has nonzero curl but zero divergence, with the sign on the curl depending on which way the function is circulating around a point (or, more precisely, the direction of $\nabla \times \mathbf{E}$ indicating, in a right-handed sense, both the plane and direction in which $\mathbf{E}$ is circulating around the point). Finally, fig. (2.16) shows an example of a vector function that has both zero divergence and zero curl: as much of the function will "go into" any closed surface as will "come out of" it, nor will there be any *net* circulation of the function one way or another around any closed loop.

2.6.1 An Important Result

To get right to the punch line: the line integral of a vector function $\mathbf{F}$ from location $\mathbf{r}_1$ to location $\mathbf{r}_2$ is independent of the path taken between those two points if and only if $\nabla \times \mathbf{F} = 0$. In addition to applications in electromagnetism, this turns out to be the criterion that determines whether it is possible to define a potential energy corresponding to a force $\mathbf{F}$.

First we prove that path independence of the line integral of $\mathbf{F}$ implies $\nabla \times \mathbf{F} = 0$. Consider the two arbitrary paths P_a and P_b from $\mathbf{r}_1$ to $\mathbf{r}_2$ shown in fig. (2.17). Since

$$\int_{P_a}^{\mathbf{r}_2} \mathbf{F} \cdot d\mathbf{r} = \int_{P_b}^{\mathbf{r}_2} \mathbf{F} \cdot d\mathbf{r}$$

we have

$$\begin{aligned} 0 &= \int_{P_a}^{\mathbf{r}_2} \mathbf{F} \cdot d\mathbf{r} - \int_{P_b}^{\mathbf{r}_2} \mathbf{F} \cdot d\mathbf{r} \\ &= \int_{P_a}^{\mathbf{r}_2} \mathbf{F} \cdot d\mathbf{r} + \int_{P_b}^{\mathbf{r}_1} \mathbf{F} \cdot d\mathbf{r} \\ &= \oint_C \mathbf{F} \cdot d\mathbf{r} \end{aligned}$$

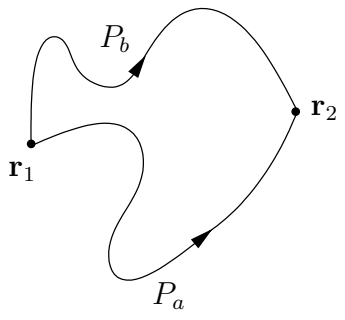


Figure 2.17: Two Paths From $\mathbf{r}_1$ to $\mathbf{r}_2$

where C is the loop formed by joining together P_a with the reverse of P_b, that is, by going from $\mathbf{r}_1$ to $\mathbf{r}_2$ along P_a and then from $\mathbf{r}_2$ back to $\mathbf{r}_1$ along P_b. Since P_a and P_b were arbitrary — they could be any paths between any pair of points we want — , we must have

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = 0$$

around all possible closed loops C. By Stokes's theorem, this means that

$$\int_S dA\, \widehat{\mathbf{n}} \cdot \nabla \times \mathbf{F} = 0 \tag{2.9}$$

for all possible surfaces S spanning all possible loops C, which can only be the case if $\nabla \times \mathbf{F} = 0$.[11]

Now we need to prove also the reverse: that if $\nabla \times \mathbf{F} = 0$, the line integral of $\mathbf{F}$ between any two points $\mathbf{r}_1$ and $\mathbf{r}_2$ is independent of the path taken between them. Using $\nabla \times \mathbf{F} = 0$ in Stokes's theorem yields

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = 0$$

for all possible loops C. If we then break these integrals around closed loops into pairs of integrals between points (exactly the reverse of what we did in the preceding paragraph), we then arrive at the conclusion that

$$\int_{P_a}^{\mathbf{r}_2} \mathbf{F} \cdot d\mathbf{r} = \int_{P_b}^{\mathbf{r}_2} \mathbf{F} \cdot d\mathbf{r}$$

for any two paths P_a and P_b between any two points $\mathbf{r}_1$ and $\mathbf{r}_2$.

2.7 A Few More Important Results

2.7.1 $\nabla \cdot \mathbf{B} = 0 \Rightarrow \mathbf{B} = \nabla \times \mathbf{A}$

In this section we establish that if $\nabla \cdot \mathbf{B} = 0$, then $\mathbf{B}$ is a pure curl, that is, there must exist some vector function $\mathbf{A}$ such that $\mathbf{B} = \nabla \times \mathbf{A}$. We will do this by explicitly constructing the $\mathbf{A}$ that will yield $\mathbf{B}$ when we take its curl. If we expand out both sides of $\mathbf{B} = \nabla \times \mathbf{A}$, we see that we need

$$\begin{aligned} &B_x(x,y,z)\,\widehat{\mathbf{x}} + B_y(x,y,z)\,\widehat{\mathbf{y}} + B_z(x,y,z)\,\widehat{\mathbf{z}} \\ &\quad = \left(\frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z}\right)\widehat{\mathbf{x}} - \left(\frac{\partial A_z}{\partial x} - \frac{\partial A_x}{\partial z}\right)\widehat{\mathbf{y}} + \left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)\widehat{\mathbf{z}} \end{aligned}$$

[11] A conclusion that would not be justified if eq. (2.9) held only for some particular surface or loop or some limited set of surfaces or loops. The reasoning is that if $\nabla \times \mathbf{F}$ were nonzero somewhere, then we could always construct some loop and surface for which $\int_S dA\, \widehat{\mathbf{n}} \cdot \nabla \times \mathbf{F} \neq 0$, which would contradict our premise that $\int_S dA\, \widehat{\mathbf{n}} \cdot \nabla \times \mathbf{F} = 0$ for all surfaces and loops.

or, in other words,

$$B_x(x,y,z) = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \tag{2.10a}$$

$$B_y(x,y,z) = -\frac{\partial A_z}{\partial x} + \frac{\partial A_x}{\partial z} \tag{2.10b}$$

$$B_z(x,y,z) = \frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y} \tag{2.10c}$$

Although it might seem like we are making our task just that much more difficult, let us impose the restriction $A_x = 0$. Then eqq. (2.10) reduce to

$$B_x(x,y,z) = \frac{\partial A_z}{\partial y} - \frac{\partial A_y}{\partial z} \tag{2.11a}$$

$$B_y(x,y,z) = -\frac{\partial A_z}{\partial x} \tag{2.11b}$$

$$B_z(x,y,z) = \frac{\partial A_y}{\partial x} \tag{2.11c}$$

To obtain A_y and A_z, all we have to do is integrate eqq. (2.11c) and (2.11b), but in doing so we have to remember that these equations are for the *partial* derivatives $\partial/\partial x$ and that we can therefore have a constant of integration that is a function of the "non-x" variables y and z:

$$A_y = \int_{x_0}^{x} dx\, B_z(x,y,z) + f(y,z) \tag{2.12a}$$

$$A_z = -\int_{x_0}^{x} dx\, B_y(x,y,z) + g(y,z) \tag{2.12b}$$

where f and g are as yet undetermined functions of y and z and where the lower limit x_0 is arbitrarily chosen.[12] As you can see by acting on eqq. (2.12a) and (2.12b) with $\partial/\partial x$, eqq. (2.11c) and (2.11b) are now satisfied; it remains only to ensure that eq. (2.11a) is also satisfied. If we substitute eqq. (2.12a) and (2.12b) into eq. (2.11a), we have

$$\begin{aligned} B_x(x,y,z) &= \frac{\partial}{\partial y}\Big(-\int_{x_0}^{x} dx\, B_y(x,y,z) + g(y,z)\Big) \\ &\qquad - \frac{\partial}{\partial z}\Big(\int_{x_0}^{x} dx\, B_z(x,y,z) + f(y,z)\Big) \\ &= \left(-\int_{x_0}^{x} dx\, \frac{\partial B_y}{\partial y} + \frac{\partial g}{\partial y}\right) - \left(\int_{x_0}^{x} dx\, \frac{\partial B_z}{\partial z} + \frac{\partial f}{\partial z}\right) \\ &= -\int_{x_0}^{x} dx \left(\frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z}\right) + \frac{\partial g}{\partial y} - \frac{\partial f}{\partial z} \end{aligned} \tag{2.13}$$

[12] Dude: that means we can choose x_0 to have whatever value we want.

To make further progress with right-hand side, we need to carry out the integration, and it very conveniently just happens that

$$\nabla \cdot \mathbf{B} = \frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} = 0$$

allows us to do so by making the integrand a total differential: using

$$\frac{\partial B_x}{\partial x} = -\left(\frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z}\right)$$

in eq. (2.13), we have

$$\begin{aligned} B_x(x,y,z) &= \int_{x_0}^{x} dx\, \frac{\partial B_x}{\partial x} + \frac{\partial g}{\partial y} - \frac{\partial f}{\partial z} \\ &= B_x(x,y,z)\Big|_{x_0}^{x} + \frac{\partial g}{\partial y} - \frac{\partial f}{\partial z} \\ &= B_x(x,y,z) - B_x(x_0,y,z) + \frac{\partial g}{\partial y} - \frac{\partial f}{\partial z} \end{aligned}$$

We therefore need f and g to be such that

$$-B_x(x_0,y,z) + \frac{\partial g}{\partial y} - \frac{\partial f}{\partial z} = 0 \tag{2.14}$$

One choice that will satisfy this requirement is

$$f = 0 \qquad\qquad g = \int_{x_0}^{y} dy\, B_x(x_0,y,z)$$

as you can verify by simply substituting this f and g into eq. (2.14).

Using only that $\nabla \cdot \mathbf{B} = 0$, we have thus succeeded in finding a vector field $\mathbf{A}$ such that $\mathbf{B} = \nabla \times \mathbf{A}$.

Note, however, that this result for $\mathbf{A}$ is not unique: we arbitrarily chose to set $A_x = 0$, the value of x_0 in the lower limit on our integrations is arbitrary, and there were other ways we could have divvied up $-B_x(x_0,y,z)$ between $\partial g/\partial y$ and $\partial f/\partial z$. In particular, if we have an $\mathbf{A}$ such that $\nabla \times \mathbf{A} = \mathbf{B}$, then we can add to this $\mathbf{A}$ an arbitrary pure gradient $\nabla\Phi$: since $\nabla \times \nabla\Phi = 0$ for any Φ, $\mathbf{A} + \nabla\Phi$ will also satisfy $\nabla \times \mathbf{A} = \mathbf{B}$.

2.7.2 The Behavior of $\nabla^2 \frac{1}{r}$

The differential operator

$$\begin{aligned} \nabla^2 &= \nabla \cdot \nabla \\ &= \left(\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right) \cdot \left(\hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y} + \hat{\mathbf{z}}\frac{\partial}{\partial z}\right) \end{aligned}$$

$$= \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

is called the Laplace operator or Laplacian.

Consider now

$$\nabla^2 \frac{1}{r} = \nabla \cdot \nabla \frac{1}{r}$$

Since

$$r = |\mathbf{r}| = |x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}| = \sqrt{x^2 + y^2 + z^2}$$

the $\nabla \frac{1}{r}$ part of $\nabla \cdot \nabla \frac{1}{r}$ is

$$\begin{aligned}
\nabla \frac{1}{r} &= \left(\hat{\mathbf{x}} \frac{\partial}{\partial x} + \hat{\mathbf{y}} \frac{\partial}{\partial y} + \hat{\mathbf{z}} \frac{\partial}{\partial z} \right) \frac{1}{\sqrt{x^2 + y^2 + z^2}} \\
&= \hat{\mathbf{x}} \frac{\partial}{\partial x} \frac{1}{\sqrt{x^2 + y^2 + z^2}} + y \text{ term} + z \text{ term}
\end{aligned}$$

which, by the chain rule, works out to

$$\begin{aligned}
&= \hat{\mathbf{x}} \left(-\frac{1}{2} \frac{1}{(x^2 + y^2 + z^2)^{\frac{3}{2}}} \frac{\partial}{\partial x} (x^2 + y^2 + z^2) \right) + y \text{ term} + z \text{ term} \\
&= \hat{\mathbf{x}} \left(-\frac{1}{2} \frac{1}{(x^2 + y^2 + z^2)^{\frac{3}{2}}} 2x \right) + y \text{ term} + z \text{ term} \\
&= -\frac{x\hat{\mathbf{x}}}{(x^2 + y^2 + z^2)^{\frac{3}{2}}} + y \text{ term} + z \text{ term} \\
&= -\frac{x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}}{(x^2 + y^2 + z^2)^{\frac{3}{2}}} \\
&= -\frac{\mathbf{r}}{r^3} = -\frac{\hat{\mathbf{r}}}{r^2}
\end{aligned} \tag{2.15}$$

If now we apply Gauss's theorem to a spherical volume V of radius r centered at the origin and its enclosing spherical surface S, we have, since $\hat{\mathbf{n}} = \hat{\mathbf{r}}$ on this spherical surface,

$$\begin{aligned}
\int_V dV\, \nabla^2 \frac{1}{r} &= \int_V dV\, \nabla \cdot \nabla \frac{1}{r} \\
&= \oint_S dA\, \hat{\mathbf{n}} \cdot \nabla \frac{1}{r} \\
&= \oint_S dA\, \hat{\mathbf{r}} \cdot \left(-\frac{\hat{\mathbf{r}}}{r^2} \right) \\
&= -\oint_S dA\, \frac{1}{r^2}
\end{aligned}$$

Since the radius r is constant over the spherical surface, this reduces to

$$\int_V dV\,\nabla^2\frac{1}{r} = -\frac{1}{r^2}\oint_S dA = -\frac{1}{r^2}\,4\pi r^2 = -4\pi$$

To see what happens when the volume V over which we are integrating $\nabla^2\frac{1}{r}$ does not include the origin, we can do out the volume integral directly using eqq. (2.15):

$$\begin{aligned}
\int_V dV\,\nabla^2\frac{1}{r} &= \int_V dV\,\nabla\cdot\nabla\frac{1}{r} \\
&= \int_V dV\,\nabla\cdot\left(-\frac{\mathbf{r}}{r^3}\right) \\
&= \int_V dV\left(\hat{\mathbf{x}}\,\frac{\partial}{\partial x} + \hat{\mathbf{y}}\,\frac{\partial}{\partial y} + \hat{\mathbf{z}}\,\frac{\partial}{\partial z}\right)\cdot\left(-\frac{x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}}{(x^2+y^2+z^2)^{\frac{3}{2}}}\right) \\
&= -\int_V dV\left(\frac{\partial}{\partial x}\left(\frac{x}{(x^2+y^2+z^2)^{\frac{3}{2}}}\right) + y\text{ term} + z\text{ term}\right)
\end{aligned}$$

Since

$$\begin{aligned}
\frac{\partial}{\partial x}\left(\frac{x}{(x^2+y^2+z^2)^{\frac{3}{2}}}\right) &= \frac{1}{(x^2+y^2+z^2)^{\frac{3}{2}}}\,\frac{\partial x}{\partial x} - x\,\frac{\partial}{\partial x}\frac{1}{(x^2+y^2+z^2)^{\frac{3}{2}}} \\
&= \frac{1}{(x^2+y^2+z^2)^{\frac{3}{2}}} - x\left(-\frac{3}{2}\frac{2x}{(x^2+y^2+z^2)^{\frac{5}{2}}}\right) \\
&= \frac{1}{r^3} - \frac{3x^2}{r^5}
\end{aligned}$$

we arrive at

$$\begin{aligned}
\int_V dV\,\nabla^2\frac{1}{r} &= -\int_V dV\left(\left(\frac{1}{r^3} - \frac{3x^2}{r^5}\right) + \left(\frac{1}{r^3} - \frac{3y^2}{r^5}\right) + \left(\frac{1}{r^3} - \frac{3z^2}{r^5}\right)\right) \\
&= -\int_V dV\left(\frac{3}{r^3} - 3\,\frac{x^2+y^2+z^2}{r^5}\right) \\
&= -\int_V dV\left(\frac{3}{r^3} - 3\frac{r^2}{r^5}\right) = 0
\end{aligned}$$

It might therefore seem that we should also have gotten zero for the previous case of a spherical volume concentric with the origin. The complication, however, is that $\nabla^2\frac{1}{r}$ is very naughty at $r = 0$. For a volume that does not include the origin, this naughtiness is not a problem, and we circumvented it in the case of a spherical volume concentric with the origin by looking only at the surface integral over a spherical surface of nonzero radius.

So far we have results for the integral of $\nabla^2\frac{1}{r}$ over spherical volumes concentric with the origin (which yield -4π) and volumes of general shape that do not include the origin (which yield zero). Since we can construct a volume of general shape that includes the origin by combining a sphere concentric with the origin with other, irregularly shaped volumes that do not include the origin, we can, however, immediately conclude that the integral $\nabla^2\frac{1}{r}$ over any volume that includes the origin will yield -4π: we will get -4π from the sphere concentric with the origin and zero from each of the irregularly shaped volumes. We thus have

$$\int_V dV\, \nabla^2\frac{1}{r} = \begin{cases} -4\pi & \text{when } V \text{ includes the origin} \\ 0 & \text{when } V \text{ does not include the origin} \end{cases}$$

In other words, the behavior of $\nabla^2\frac{1}{r}$ is such that it vanishes everywhere except at the origin, where it has an infinite spike that yields -4π when integrated over. Since the only contribution to any integral that includes $\nabla^2\frac{1}{r}$ in the integrand will be from this spike at $\mathbf{r} = 0$, we will also have, for any nice function $f(\mathbf{r})$,

$$\int_V dV\, f(\mathbf{r})\, \nabla^2\frac{1}{r} = \begin{cases} -4\pi f(0) & \text{when } V \text{ includes the origin} \\ 0 & \text{when } V \text{ does not include the origin} \end{cases} \tag{2.16}$$

where $f(0)$ is the value of f at the origin.

On the face of it, this may not look like a terribly important or useful result, but appearances can be deceiving. We wouldn't be surprised if a use for eq. (2.16) arose in the very next section.

2.7.3 Helmholtz's Theorem

Helmholtz's theorem states that any vector function $\mathbf{V} = \mathbf{V}(x, y, z)$ can be resolved into the sum of a pure gradient and a pure curl, that is, that it can be written as

$$\mathbf{V} = -\nabla\varphi + \nabla \times \mathbf{A}$$

for some $\varphi = \varphi(x, y, z)$ and $\mathbf{A} = \mathbf{A}(x, y, z)$. The scalar function φ and the vector function $\mathbf{A}$ are known as potential functions (specifically, φ as the scalar potential and $\mathbf{A}$ as the vector potential). Believe it or not, we will actually need this theorem when we get to electromagnetism and want to express the electric and magnetic fields in terms of potentials.

To prove the theorem, let us suppose (as we will in fact prove below) that we can find a function φ such that

$$\nabla \cdot \mathbf{V} = -\nabla^2\varphi$$

Then we will have

$$\begin{aligned}\nabla \cdot (\mathbf{V} + \nabla\varphi) &= \nabla \cdot \mathbf{V} + \nabla \cdot \nabla\varphi \\ &= \nabla \cdot \mathbf{V} + \nabla^2\varphi \\ &= \nabla \cdot \mathbf{V} - \nabla \cdot \mathbf{V} \\ &= 0\end{aligned}$$

According to our results in §2.7.1, this means that $\mathbf{V} + \nabla\varphi$ must be pure curl, that is, must be expressible in the form

$$\mathbf{V} + \nabla\varphi = \nabla \times \mathbf{A}$$

so that

$$\mathbf{V} = -\nabla\varphi + \nabla \times \mathbf{A}$$

Proving Helmholtz's theorem therefore boils down to our being able to find a function φ such that $\nabla \cdot \mathbf{V} = -\nabla^2\varphi$.

Our search for this holy grail will unfortunately not involve Monty Python, but fortunately neither will it take long. We boldly assert that the φ we want is given by [13]

$$\varphi = \frac{1}{4\pi}\int dV' \, \frac{\psi(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \qquad \text{with} \quad \psi = \nabla \cdot \mathbf{V} \tag{2.17}$$

where the integration by $dV' = dx'\,dy'\,dz'$ is over all of space and where

$$\begin{aligned}\mathbf{r} &= x\,\widehat{\mathbf{x}} + y\,\widehat{\mathbf{y}} + z\,\widehat{\mathbf{z}} \\ \mathbf{r}' &= x'\,\widehat{\mathbf{x}} + y'\,\widehat{\mathbf{y}} + z'\,\widehat{\mathbf{z}}\end{aligned}$$

To verify this claim, we act on both sides of eq. (2.17) with ∇^2, being mindful that

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

acts on the coordinates (x, y, z) and not on the coordinates (x', y', z'):

$$\begin{aligned}\nabla^2\varphi &= \nabla^2\frac{1}{4\pi}\int dV' \, \frac{\psi(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \\ &= \frac{1}{4\pi}\int dV' \, \psi(\mathbf{r}')\nabla^2\frac{1}{|\mathbf{r} - \mathbf{r}'|}\end{aligned} \tag{2.18}$$

Now,

$$\nabla^2\frac{1}{|\mathbf{r} - \mathbf{r}'|}$$

[13] At this point, you have to imagine an orchestra playing Copland's *Fanfare for the Common Man*.

is just like $\nabla^2 \frac{1}{r}$, except that instead of having a spike at $r = 0$, at has a spike at $|\mathbf{r} - \mathbf{r}'| = 0$, that is, at $\mathbf{r}' = \mathbf{r}$. So according to our result (2.16) from the preceding section, integrating over $\nabla^2(1/|\mathbf{r} - \mathbf{r}'|)$ will yield zero everywhere except at the point $\mathbf{r}' = \mathbf{r}$, where the spike in $\nabla^2(1/|\mathbf{r} - \mathbf{r}'|)$ will give a -4π. Eq. (2.18) thus simplifies to

$$\nabla^2\varphi = \frac{1}{4\pi}\Big[-4\pi\psi(\mathbf{r}')\Big]_{\mathbf{r}'=\mathbf{r}} = \psi(\mathbf{r})$$

Since ψ was defined as $\nabla \cdot \mathbf{V}$, we have what we need to complete the proof: $\nabla \cdot \mathbf{V} = -\nabla^2\psi$. Word.

2.8 Problems

1. Evaluate
$$\int_S dA\,(x^2+y^2)$$
where S is a rectangle aligned with the x and y axes and extending from $(-a,-b)$ to (a,b).

2. Evaluate
$$\int_S dA\,r^2$$
where S is the semicircle of radius a extending from $(0,-a)$ to $(0,a)$.

3. Evaluate
$$\int_S dA\,r^2$$
where S is the first octant of the spherical surface $r=a$. (First octant meaning the octant in which x, y, and z are all positive.)

4. Evaluate
$$\int_V dV\,(x^2+y^2)$$
where V is a rectangular box aligned with the x, y, and z axes and extending from $(-a,-b,-c)$ to (a,b,c).

5. Evaluate
$$\int_V dV\,r^2$$
where V is the cylndrical volume of radius $r=a$ extending from $z=-b$ to $z=b$.

6. Evaluate
$$\int_V dV\,r^2$$
where V is the volume of a sphere of radius $r=a$.

7. Evaluate the line integral of the vector function
$$\mathbf{F} = y\,\hat{\mathbf{x}} - x\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}$$
along the straight line segment from the origin to (a,a,a). See the footnote if you need a hint.[14]

8. Evaluate the line integral of the vector function
$$\mathbf{F} = F_0\,\widehat{\boldsymbol{\theta}} \qquad (F_0 = \text{const})$$
along the circular arc from $(a,0)$ to $(0,a)$.

[14]Note that along this line segment $x=y=z$.

9. (a) Evaluate the line integral of the vector function

$$\mathbf{F} = y\,\hat{\mathbf{x}}$$

counterclockwise around the circle $x^2 + y^2 = a^2$. See the footnote if you need a hint.[15]

(b) Show that you get, as Stokes's theorem dictates, the same result for

$$\int_S dA\,\hat{\mathbf{n}} \cdot \nabla \times \mathbf{F}$$

where S is the planar surface spanning the circle $x^2 + y^2 = a^2$.

10. (a) Evaluate the line integral of the vector function

$$\mathbf{F} = y\,\hat{\mathbf{x}}$$

counterclockwise around a square of side a that lies in the xy plane and is centered at the origin with edges aligned with the coordinate axes. See the footnote if you need a hint.[16]

(b) Show that you get, as Stokes's theorem dictates, the same result for

$$\int_S dA\,\hat{\mathbf{n}} \cdot \nabla \times \mathbf{F}$$

where S is the planar surface spanning the square described in the previous part.

11. (a) Show that the vector function

$$\mathbf{F} = \frac{k}{r^2}\,\hat{\mathbf{r}} = \frac{k}{r^3}\,\mathbf{r}$$

where k is a positive constant, is conservative. See the footnote if you need a hint.[17]

[15] To integrate around this circle, you will want to use the polar or cylindrical expression for $d\mathbf{r}$ and to remember that the radius is constant ($r = a$).

[16] You need to break the line integral into four pieces, one along each side of the square. And along the sides you have either $x = \pm\frac{1}{2}a$ with $dx = 0$, or $y = \pm\frac{1}{2}a$ with $dy = 0$.

[17] The only expression we have for the curl is in Cartesian coordinates, so you need to express r as $x^2 + y^2 + z^2$ and $\mathbf{r}$ as $x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}$. Be sure not to do unnecessary work when you actually evaluate the curl.

(b) More generally, show that any vector function that depends only on the distance r from the origin and that is either radially inward toward the origin or radially outward from it — that is, any vector function that can be written in the form

$$\mathbf{F} = f(r)\,\mathbf{r}$$

— is conservative.

When we get to dynamics and energy, this will have the important consequence that any central force, such as Newton's law of gravity, is conservative and can therefore be expressed as the gradient of a potential energy.

12. Show that Gauss's theorem holds for the vector function

$$\mathbf{F} = -kx\,\hat{\mathbf{x}}$$

where k is a positive constant, by explicitly evaluating

$$\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{F} \qquad \text{and} \qquad \int_V dV\,\nabla\cdot\mathbf{F}$$

where S and V are the surface and volume of a cube of side a, centered at the origin, with edges aligned with the coordinate axes. See the footnote if you need a hint.[18]

13. Use

$$\Phi = \int_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E}$$

to calculate the flux Φ of the electric field

$$\mathbf{E} = \frac{2\lambda}{r}\,\hat{\mathbf{r}} \qquad (\lambda = \text{const})$$

through the side of a cylindrical surface of radius $r = a$ and length ℓ. (By *side* we mean the cylindrical surface without the end-caps.)

14. Use

$$\Phi = \int_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E}$$

to calculate the flux Φ of the electric field

$$\mathbf{E} = \frac{q}{r^2}\,\hat{\mathbf{r}} \qquad (q = \text{const})$$

through the through a spherical surface of radius $r = a$.

[18] You need to break the surface integral into six pieces, one over each face of the cube, with $\hat{\mathbf{n}} = \pm\hat{\mathbf{x}}$ for the $x = \pm\frac{1}{2}a$ faces, etc.

15. Suppose you find yourself in a bizarre terrain where your height (altitude) is given by

$$h = \alpha xy$$

where α is a positive constant.

(a) What are the physical dimensions of α?

(b) Evaluate ∇h.

(c) If you are at the point $(3a, 4a)$, where a is some given distance, what direction is most steeply uphill? Specify the direction by giving a unit vector that points in that direction.

(d) If you start to travel in that direction from the point $(3a, 4a)$, at what rate is your altitude changing?

(e) What are the dimensions of your answer to the preceding part?

(f) At the point $(3a, 4a)$, in what directions could you start to move without changing your altitude? Specify these directions by giving counterclockwise angles with the positive x axis.

2.9 Sketchy Answers

(1) $\frac{4}{3}ab(a^2+b^2)$.

(2) $\frac{1}{4}\pi a^4$.

(3) $\frac{1}{2}\pi a^4$.

(4) $\frac{8}{3}abc(a^2+b^2)$.

(5) $\pi a^4 b$.

(6) $\frac{4}{5}\pi a^5$.

(7) $\frac{1}{2}a^2$.

(8) $\frac{\pi a}{2}F_0$.

(9) $-\pi a^2$.

(10) $-a^2$.

(12) $-ka^3$.

(13) $4\pi\lambda\ell$.

(14) $4\pi q$.

(15b) $\alpha(y\,\widehat{\mathbf{x}} + x\,\widehat{\mathbf{y}})$.

(15c) $\frac{4}{5}\,\widehat{\mathbf{x}} + \frac{3}{5}\,\widehat{\mathbf{y}}$.

(15d) $5\alpha a$.

(15f) $-\tan^{-1}\frac{4}{3}$ and $\pi - \tan^{-1}\frac{4}{3}$.

Part II

Basic Mechanics

Chapter 3

Kinematics

3.1 Location, Velocity, & Acceleration

In physics, the term mechanics covers all the physics related to the motion of bodies. Mechanics begins with kinematics, which describes motion quantitatively (without concerning itself with what causes motion).

In Newtonian mechanics — the physics that to a good approximation accounts for the motion of bodies in everyday situations[1] — the motion of a body can be described by just a three basic parameters:[2] location, velocity, and acceleration.

Specifying the location of a body tells you (*duh!*) where it is. To specify location quantitatively in three-dimensional space, we can set up an xyz coordinate system; the values of the three coordinates x, y, and z then tell us where to find the body in each of the three spatial directions.[3] Location is a therefore a vector quantity, in the sense that we can draw a vector from the origin of our coordinate system to the point (x, y, z) where the body is located: it is not enough to say just how far the body is from the origin; we must also specify the direction. The vector that goes from the origin to the body is called the body's position vector and is conventionally denoted by $\mathbf{r}$.[4] Since $\mathbf{r}$ extends from the origin to the point (x, y, z), the components of $\mathbf{r}$ are simply $r_x = x$, $r_y = y$, and $r_z = z$, as shown for the two-dimensional

[1] Examples of situations not to be considered "everyday": you are moving close to the speed of light, you are the size of an atom, you are near a black hole, you have a single large eye in the middle of your forehead, etc.

[2] Actually, this is ignoring any rotational motion, which we will deal with later.

[3] For the most part, we will, however, be working with motion in just one or two dimensions, so that we need only an x axis or an xy plane. But we will keep the present discussion general to three dimensions.

[4] This notation is used because $\mathbf{r}$, since it starts from the origin, could also be thought of as a radius.

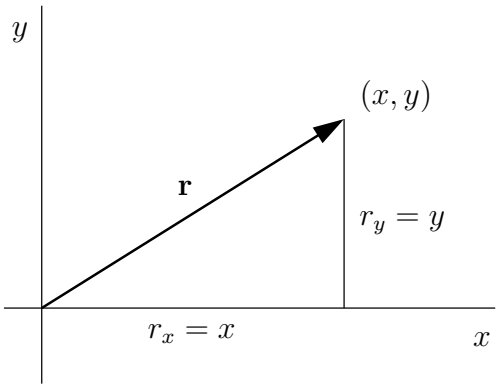


Figure 3.1: A Two-Dimensional Position Vector

case in fig. (3.1):

$$\mathbf{r} = x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}$$

If a body is moving, this location will be a function of the time, t:

$$\mathbf{r}(t) = x(t)\,\hat{\mathbf{x}} + y(t)\,\hat{\mathbf{y}} + z(t)\,\hat{\mathbf{z}} \tag{3.1}$$

Note that it is only the coordinates (x, y, z) of the body's location that change; the unit vectors $\hat{\mathbf{x}}$, $\hat{\mathbf{y}}$, and $\hat{\mathbf{z}}$ are constant: they always have unit length, and they always point down their respective positive axes.

Displacement is defined as the difference between a body's final and initial locations. Since the initial and final locations are vector quantities, the displacement is also a vector: if the body goes from location $\mathbf{r}_i$ to location $\mathbf{r}_f$, then its vector displacement $\Delta\mathbf{r}$ is

$$\Delta\mathbf{r} = \mathbf{r}_f - \mathbf{r}_i$$

as shown in fig. (3.2).

Be careful to distinguish displacement from the *distance* traveled by the body: while displacement is a vector, distance does not take into account direction and is therefore just a scalar. Also, the distance traveled depends not only on the initial and final locations, but also on the path taken between

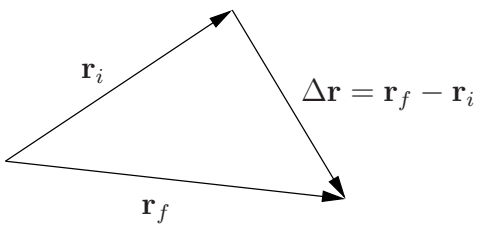


Figure 3.2: A Displacement Vector

them. If, for example, you take one step to the right, then one step to the left, then one step north, your overall displacement is just one step north, even though, for the path you took, you traveled a distance of three steps.

Velocity measures how rapidly and in what direction the location of the body is changing, and, like displacement, is a vector quantity. Specifically, velocity is defined as the rate of change of a body's location — that is, its displacement divided by the corresponding time interval:

$$\mathbf{v} = \frac{\Delta \mathbf{r}}{\Delta t} \tag{3.2}$$

A body's velocity may, of course, vary with time, in which case eq. (3.2) will give only the average velocity over the finite time interval from t to $t + \Delta t$. But if we average over smaller and smaller time intervals Δt, this average velocity will more and more closely approximate the body's velocity exactly at time t, and in the limit that Δt goes to zero, it will in fact give us the body's instantaneous velocity at time t:

$$\mathbf{v} = \lim_{\Delta t \to 0} \frac{\Delta \mathbf{r}}{\Delta t} = \frac{d\mathbf{r}}{dt} \tag{3.3}$$

So if we know the body's location as a function of time (see eq. (3.1)), we can obtain its instantaneous velocity as a function of time simply by taking the time derivative: [5]

$$\begin{aligned} \mathbf{v}(t) &= \frac{d\mathbf{r}}{dt} \\ &= \frac{d}{dt}\Big(x(t)\,\hat{\mathbf{x}} + y(t)\,\hat{\mathbf{y}} + z(t)\,\hat{\mathbf{z}}\Big) \\ &= \frac{dx}{dt}\,\hat{\mathbf{x}} + \frac{dy}{dt}\,\hat{\mathbf{y}} + \frac{dz}{dt}\,\hat{\mathbf{z}} \end{aligned} \tag{3.4}$$

Thus the components of the velocity are

$$v_x = \frac{dx}{dt} \qquad v_y = \frac{dy}{dt} \qquad v_z = \frac{dz}{dt} \tag{3.5}$$

Geometrically, the derivative of a function is tangent to that function's curve. Eq. (3.3) therefore tells us that at any given moment, a body's velocity is tangent to the curve traced out by the time evolution of its position vector $\mathbf{r}$ — that is, a body's velocity is tangent to its path (trajectory). Fig. (3.3) illustrates this for three points on a body's path P. [6]

[5] Again, we have noted that the unit vectors are all constant and that therefore the derivative acts only on the x, y, and z.

[6] Fig. (3.3) is meant to illustrate only that the *direction* of the velocity is tangent to the path of the body; no inference should be drawn about the *magnitude* of the velocity, since it is possible to follow the path shown in the figure (or indeed any path, for that matter) at any speed. The magnitudes of the velocity vectors in the figure have been chosen randomly.

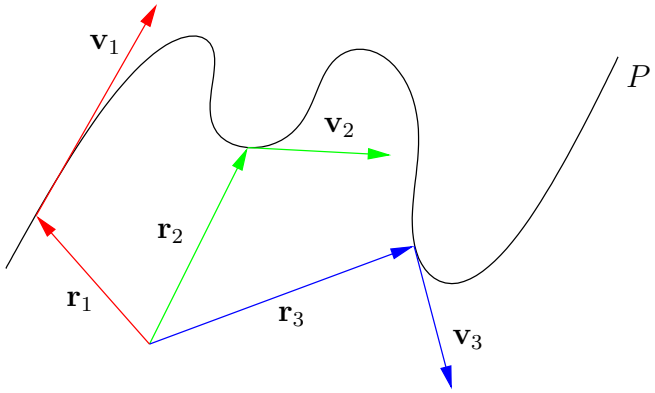


Figure 3.3: Yo: Velocity is Tangent to the Path

Just as displacement is a vector and must be distinguished from the scalar distance, velocity is a vector and must be distinguished from the scalar speed: speed is just the magnitude of the velocity; it measures simply how fast the body is moving, without taking into account direction. If, for example, you suddenly get up, run around the room screaming "Narf! Narf! Narf!" at a steady 8 mph, and then return to your seat, your average speed is 8 mph, but your average velocity is zero because you have returned to your starting point and your net displacement is therefore zero.

Location and velocity are fairly intuitive: it is not hard to visualize, say, a lacrosse ball being 1 ft in front of your face and coming directly toward you at 100 mph. Acceleration, however, is more abstract: it measures the rate of change velocity. Since velocity is a vector, there are two ways a body can be accelerating: the magnitude of the body's velocity (its speed) can be changing — that is, the body can be speeding up or slowing down — , or, even if the speed remains constant, the direction of the body's motion can be changing. If, for example, you do donuts in the parking lot at a steady speed of 10 mph, the direction of your velocity (the direction in which you are headed) is continually changing as you circle around, so that you are experiencing an acceleration even though your speed is a constant 10 mph. Moreover, the acceleration itself is a vector quantity: its magnitude measures how rapidly the body's velocity is changing, and — as we will show at the end of this section — its direction indicates in which of the above two ways that change is occurring: to the extent that the acceleration is parallel to the velocity, the body is speeding up or slowing down, while to the extent that the acceleration is perpendicular to the velocity, the body is changing its direction of motion.

Though acceleration may be more difficult to get an intuitive feel for, the analytic relations for it are very straightforward: the definition of acceleration as the rate of change of velocity is exactly analogous to the definition of

velocity as the rate of change of displacement. By analogy to eqq. (3.2), (3.3), and (3.4), we therefore have for the average acceleration

$$\mathbf{a} = \frac{\Delta \mathbf{v}}{\Delta t} \tag{3.6}$$

and for the instantaneous acceleration

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} \tag{3.7}$$

and

$$\begin{aligned} \mathbf{a}(t) &= \frac{d\mathbf{v}}{dt} \\ &= \frac{d}{dt}\Big(v_x(t)\,\hat{\mathbf{x}} + v_y(t)\,\hat{\mathbf{y}} + v_z(t)\,\hat{\mathbf{z}}\Big) \\ &= \frac{dv_x}{dt}\,\hat{\mathbf{x}} + \frac{dv_y}{dt}\,\hat{\mathbf{y}} + \frac{dv_z}{dt}\,\hat{\mathbf{z}} \end{aligned} \tag{3.8}$$

Thus the components of the acceleration are

$$a_x = \frac{dv_x}{dt} \qquad a_y = \frac{dv_y}{dt} \qquad a_z = \frac{dv_z}{dt} \tag{3.9}$$

Now that we've defined location, velocity, and acceleration analytically, a few general observations about them:

First, it would of course be possible to continue the above series of definitions: if velocity is the rate of change of location and acceleration is the rate of change of velocity, why not go on to define something like "hyper-acceleration" as the rate of change of the acceleration, and so on? Since kinematics proper is concerned only with describing motion and not with its causes, there is no reason why it shouldn't deal with hyper-acceleration or even the whole infinite series of analogous quantities defined by taking rates of change. But when we get to dynamics, which does concern itself with the causes of motion, we will see that it is forces that give rise to motion, and forces turn out to be related to acceleration. Although such animals could be defined, there is therefore no need for the purposes of physics to deal with any rates of change beyond acceleration.

Second, note that in eqq. (3.4) and (3.8) the location, velocity, and acceleration all separate by component: the x component of the velocity depends only on the x coordinate ($v_x = dx/dt$) and not on the y or z coordinates, and the x component of the acceleration depends only on the x component of the velocity ($a_x = dv_x/dt$) and not on its y or z components. And likewise for the y and z components.[7] So although the motions of a body in the x, y, and z

[7] In less precise but perhaps more readily comprehensible words: the x stuff depends only on x stuff, the y stuff only on y stuff, and the z stuff only on z stuff.

directions are coupled by their common parametrization by the time t (that is, although $x = x(t)$, $y = y(t)$, and $z = z(t)$), eqq. (3.4) and (3.8) tell us that we can calculate the motions in these three spatial directions separately. This will make analyzing the motion much easier in some cases, as we will first see when we get to projectile motion.

Finally, we can combine eqq. (3.5) and (3.8) to relate acceleration directly to the second derivatives of the coordinates:

$$\begin{aligned}\mathbf{a}(t) &= \frac{dv_x}{dt}\,\hat{\mathbf{x}} + \frac{dv_y}{dt}\,\hat{\mathbf{y}} + \frac{dv_z}{dt}\,\hat{\mathbf{z}} \\ &= \frac{d}{dt}\left(\frac{dx}{dt}\right)\hat{\mathbf{x}} + \frac{d}{dt}\left(\frac{dy}{dt}\right)\hat{\mathbf{y}} + \frac{d}{dt}\left(\frac{dz}{dt}\right)\hat{\mathbf{z}} \\ &= \frac{d^2x}{dt^2}\,\hat{\mathbf{x}} + \frac{d^2y}{dt^2}\,\hat{\mathbf{y}} + \frac{d^2z}{dt^2}\,\hat{\mathbf{z}} \end{aligned} \tag{3.10}$$

Thus the components of the acceleration can also be expressed as

$$a_x = \frac{d^2x}{dt^2} \qquad a_y = \frac{d^2y}{dt^2} \qquad a_z = \frac{d^2z}{dt^2} \tag{3.11}$$

Now that we have laid out the relations defining acceleration, let us return to the question of its vector interpretation. We would like to relate the change in the body's speed $v = |\mathbf{v}|$ to the relative directions of its acceleration $\mathbf{a} = d\mathbf{v}/dt$ and its velocity $\mathbf{v}$. The "trick" is to recognize that expressing the speed v in terms of the dot product of the vector velocity $\mathbf{v}$ with itself will allow us to relate the change dv in speed to the relative directions of the velocity $\mathbf{v}$ and the change $d\mathbf{v}$ in the velocity (and hence, by $d\mathbf{v} = \mathbf{a}\,dt$, the acceleration):

$$v = \sqrt{v^2} = \sqrt{\mathbf{v}\cdot\mathbf{v}}$$

so that

$$dv = d\sqrt{\mathbf{v}\cdot\mathbf{v}}$$

Using the chain and product rules to evaluate the differential on the right-hand side, and reverting from $\sqrt{\mathbf{v}\cdot\mathbf{v}}$ back to v, we have[8]

$$dv = \frac{1}{2\sqrt{\mathbf{v}\cdot\mathbf{v}}}\,d(\mathbf{v}\cdot\mathbf{v})$$

[8] That the product rule can be applied to a dot product is easily seen using its analytic definition: in three dimensions,

$$\begin{aligned} d(\mathbf{a}\cdot\mathbf{b}) &= d(a_xb_x + a_yb_y + a_zb_z) \\ &= (da_x)b_x + (da_y)b_y + (da_z)b_z + a_x(db_x) + a_y(db_y) + a_z(db_z) \\ &= d\mathbf{a}\cdot\mathbf{b} + \mathbf{a}\cdot d\mathbf{b} \end{aligned}$$

$$= \frac{1}{2v}(d\mathbf{v}\cdot\mathbf{v} + \mathbf{v}\cdot d\mathbf{v})$$
$$= \frac{1}{2v}\,2\mathbf{v}\cdot d\mathbf{v}$$
$$= \frac{1}{v}\mathbf{v}\cdot\mathbf{a}\,dt$$

where we have in the last line used $d\mathbf{v} = \mathbf{a}\,dt$. We can now see what happens when, holding everything else fixed, we change the angle between the velocity vector $\mathbf{v}$ and the acceleration vector $\mathbf{a}$: To the extent that the velocity and acceleration are in the same direction, the dot product on the right-hand side is positive and we get an increase in speed ($dv > 0$), with this increase in speed being maximal when the velocity and acceleration are aligned. To the extent that the velocity and acceleration are in opposite directions, the dot product on the right-hand side is negative and we get a decrease in speed ($dv < 0$), with this decrease in speed being extremal when the velocity and acceleration are anti-aligned. And to the extent that the velocity and acceleration are perpendicular to each other, the dot product on the right-hand side vanishes and we get no change in speed ($dv = 0$). Since this means that there is no change in the magnitude of the velocity vector, we conclude that an acceleration perpendicular to the velocity corresponds to a change only in the direction of motion.

3.2 One-Dimensional Motion

For the one-dimensional case of a body moving along a straight line, we need only x, v_x, and a_x to describe the body's motion:

$$v_x = \frac{dx}{dt} \qquad\qquad a_x = \frac{dv_x}{dt} = \frac{d^2x}{dt^2}$$

or, for average velocity and acceleration,

$$v_x = \frac{\Delta x}{\Delta t} \qquad\qquad a_x = \frac{\Delta v_x}{\Delta t}$$

Since we only deal in this case with x components, we can omit the x subscripts for simplicity and write

$$v = \frac{dx}{dt} \qquad\qquad a = \frac{dv}{dt} = \frac{d^2x}{dt^2} \tag{3.12a}$$

$$v_{avg} = \frac{\Delta x}{\Delta t} \qquad\qquad a_{avg} = \frac{\Delta v}{\Delta t} \tag{3.12b}$$

In this one-dimensional case, directions are indicated simply by sign: a positive velocity indicates motion in the positive x direction, a negative velocity motion in the negative x direction. The physical interpretation of the

direction of the acceleration is, however, a bit more involved than that of the velocity: as we go from time t to time $t + dt$, the velocity is going from v to $v + dv$, and, as you can see from $a = dv/dt$, the sign on the acceleration corresponds to the sign on the *change* dv in velocity. We must therefore consider the two possible signs on the acceleration (that is, on dv) in conjunction with the two possible signs on the velocity v itself — a total of four cases:

$++$ For the case of positive a (positive dv) and positive v, the final velocity $v + dv$ will be a larger positive number than v — that is, the body is speeding up as it moves in the positive x direction.

$--$ Similarly, for the case of negative a (negative dv) and negative v, we get a final velocity $v + dv$ that is a "larger" negative number than v — that is, the body is speeding up as it moves in the negative x direction.

$+-$ For the case of negative a (negative dv) and positive v, the final velocity $v + dv$ will be a smaller positive number than v — that is, the body is slowing down as it moves in the positive x direction.

$-+$ And similarly for the case of positive a (positive dv) and negative v, the final velocity $v + dv$ will be a "smaller" negative number than v — that is, the body is slowing down as it moves in the negative x direction.

All of these cases may be summarized as

> When the velocity and acceleration have the same sign, the body is speeding up; when they have opposite signs, the body is slowing down.

Since geometrically the derivative of a function is the slope of that function's curve, the velocity $v = dx/dt$ is the slope of the plot of location x versus time t (that is, the plot of $x(t)$). Likewise the acceleration $a = dv/dt$ is the slope of the plot of velocity v versus time t (the plot of $v(t)$).

Relations (3.12) allow us to calculate the velocity $v(t)$ if we know the location $x(t)$ of the body as a function of time t and to calculate the acceleration $a(t)$ of the body if we know its velocity $v(t)$ as a function of time t. Since the inverse of differentiation is integration,[9] we can also do the reverse and calculate $x(t)$ from $v(t)$ and $v(t)$ from $a(t)$. To get x from v, all we have to do is rewrite $v = dx/dt$ as $dx = v\,dt$ and integrate:

$$\int_{x_i}^{x_f} dx = \int_{t_i}^{t_f} v\,dt$$

[9]Hence the cool alternative term "antidifferentiation" for it.

Note that the limits on each side of the equation correspond, as they must, to the variable being integrated over: t on the right-hand side and x on the left-hand side. Our choice of initial time t_i and final time t_f for the lower and upper limits on right-hand side are arbitrary — they can represent any times we want — , but note that we have been careful to make the lower and upper limits on the left-hand side consistent: x_i represents the location of the body at time t_i and x_f its location at time t_f. To carry out the integration on the right-hand side we would have to know v explicitly as a function of time, but the dx on the left-hand side is a total differential. Integrating the left-hand side and putting in the limits, we thus have

$$x_f - x_i = \int_{t_i}^{t_f} v\,dt \tag{3.13}$$

Note that what eq. (3.13) directly tells us is the *displacement* $\Delta x = x_f - x_i$ of the body between times t_i and t_f; to determine the absolute location x_f of the body, we would need to know its initial location x_i.[10]

Frequently we will be making the special choice $t_i = 0$ for the initial time. In this case, it is convenient to denote the initial values of quantities by a subscript 0 and their final values simply by omitting any subscript. Thus our initial and final times will be denoted by $t_0 = 0$ and t and our initial and final velocities by v_0 and v, and eq. (3.13) will become

$$x - x_0 = \int_0^t v\,dt \tag{3.14}$$

Comparing $a = dv/dt$ to $v = dx/dt$, we see that we will get relations between v and a analogous to those between x and v. Thus eqq. (3.13) and (3.14) become

$$v_f - v_i = \int_{t_i}^{t_f} a\,dt \tag{3.15}$$

$$v - v_0 = \int_0^t a\,dt \tag{3.16}$$

Since geometrically the integral of a function is the area under that function's curve, eq. (3.13) is telling us that the displacement $\Delta x = x_f - x_i$ is the area under the plot of velocity v versus time t (the plot of $v(t)$). Likewise eq. (3.15) is telling us that the change $\Delta v = v_f - v_i$ in velocity is the area under the plot of acceleration a versus time t (the plot of $a(t)$).

[10] Sometimes math courses leave people with the bad habit of making every integration indefinite and just sticking in a constant of integration. While it is possible to do this — the constant of integration can be fitted to the boundary conditions specified by the limits on the integrations — , it is very awkward and obscures the physics. It is far better to put explicit limits on your integrations. So get in the habit of using limits.

When dealing with relative velocities, just use your intuition and common sense — trying to rely on a general relation for relative velocities would lead you astray more often than it would help. If, for example, you can toss a lacrosse ball at 100 mph, and you toss one at your friend's face while running toward him or her at 20 mph, then the ball will be traveling at 120 mph relative to (that is, as perceived by) your friend.

3.2.1 Constant Acceleration

The integration in eq. (3.16) cannot be carried out unless we know the acceleration a as a function of time. It turns out, however, that the special case of constant acceleration occurs fairly frequently in practice.[11] There are three ways to work out this special case. First, we can work analytically from eq. (3.16):[12]

$$v - v_0 = \int_0^t a\, dt = a \int_0^t dt = at\Big|_0^t = at$$

or in other words

$$v = v_0 + at \tag{3.17}$$

Using this result in eq. (3.14), we further obtain

$$\begin{aligned} x - x_0 &= \int_0^t v\, dt \\ &= \int_0^t (v_0 + at)\, dt \\ &= \left[v_0 t + \tfrac{1}{2} a t^2\right]_0^t \\ &= v_0 t + \tfrac{1}{2} a t^2 \end{aligned} \tag{3.18}$$

The two equations (3.17) and (3.18) allow us to determine the motion of the body completely: if we know the body's (constant) acceleration a and its initial velocity v_0 and initial location x_0, we can determine where it is (its

[11] As we will see when we get to dynamics, the acceleration of a body will be constant whenever the net force on it is constant, and there will turn out to be many everyday situations for which this is true to a good approximation.

[12] Another habit people sometimes pick up in math courses is using "dummy" variables of integration. The ostensible motivation for this would, in this case, be to avoid confusion between the t of the upper limit and the dt by which we are integrating, and proponents of dummy variables would therefore instead write

$$\int_0^t a\, dt'$$

Grownups, however, don't use such dummy variables, so we won't either.

location x) and how fast and in what direction it's moving (its velocity v) at any time t, whether past or future.

All of the information for the case of constant acceleration is in the two equations eq. (3.17) and eq. (3.18), but these two equations can be combined algebraically to yield three other relations. These other three relations, being algebraically redundant with eq. (3.17) and eq. (3.18), have no new information in them, but they are frequently useful for solving problems. If we rewrite eq. (3.17) as $v_0 = v - at$, then eq. (3.18) becomes

$$x - x_0 = v_0 t + \tfrac{1}{2}at^2 = (v - at)t + \tfrac{1}{2}at^2 = vt - \tfrac{1}{2}at^2$$

so that we have the new relation

$$x - x_0 = vt - \tfrac{1}{2}at^2 \tag{3.19}$$

If now we add eqq. (3.18) and (3.19), we have

$$2(x - x_0) = (v_0 t + \tfrac{1}{2}at^2) + (vt - \tfrac{1}{2}at^2) = (v_0 + v)t$$

or in other words

$$x - x_0 = \frac{v_0 + v}{2}\, t \tag{3.20}$$

Eq. (3.20) is telling us that the body's displacement is its *average* velocity times the time. Finally, if we rewrite eq. (3.17) as $t = (v - v_0)/a$, then eq. (3.20) becomes

$$x - x_0 = \frac{v_0 + v}{2}\, t = \frac{v_0 + v}{2}\, \frac{v - v_0}{a} = \frac{v^2 - v_0^2}{2a}$$

or in other words

$$v^2 - v_0^2 = 2a(x - x_0) \tag{3.21}$$

Which of eq. (3.17)-(3.21) is useful depends on the situation we are analyzing. Suppose, for example, that when you floor it, you can go from rest at one end of the 40 m parking lot of a daycare center to 100 mph $\approx$ 45 m/s at the other end. If you want to determine your acceleration, the most convenient relation is eq. (3.21): all of the quantities in eq. (3.21) are known ($v_0 = 0$, $v = 45$ m/s, $x - x_0 = 40$ m) except the acceleration a, which is what you want to determine. All of the other four relations involve the time t, which you neither know nor care about.[13]

[13] You could, of course, use eq. (3.20) to solve for t and then get a from any of the three remaining equations, but that would be doing unnecessary work.

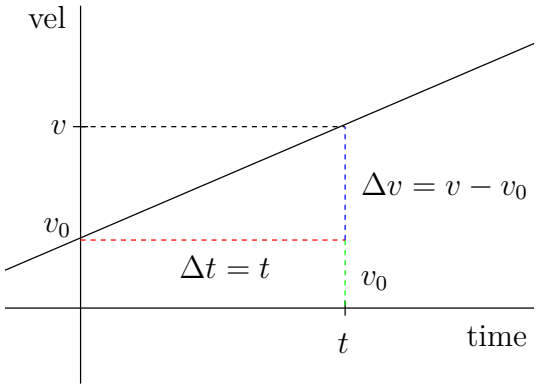


Figure 3.4: Geometry of the Velocity Plot

The second way to work out the special case of constant acceleration is geometrically: since the acceleration is the slope of the v versus t curve, we have, from fig. (3.4),

$$a = \frac{\Delta v}{\Delta t} = \frac{v - v_0}{t}$$

which, when solved for v, yields eq. (3.17). And from this same plot, we can also see that, since the displacement $x - x_0$ is the area under the v versus t curve,

$$\begin{aligned} \text{displacement} &= \text{area of triangle underneath curve} \\ &\quad + \text{ area of rectangle underneath triangle} \\ x - x_0 &= \tfrac{1}{2}\Delta v \Delta t + v_0 \Delta t \\ &= \tfrac{1}{2}(v - v_0)t + v_0 t \end{aligned}$$

which, if we use eq. (3.17) to substitute $v_0 + at$ for v, becomes

$$\begin{aligned} x - x_0 &= \tfrac{1}{2}(v_0 + at - v_0)t + v_0 t \\ &= \tfrac{1}{2}at^2 + v_0 t \end{aligned}$$

which is just eq. (3.18).

The third way to work out the special case of constant acceleration is to simply reason it out directly: since the acceleration, which is the rate of change of velocity, is constant,

$$\begin{aligned} \text{velocity} &= \text{initial velocity} + \text{change in velocity} \\ &= \text{initial velocity} + \text{rate} \times \text{time} \\ &= v_0 + at \end{aligned}$$

which is again eq. (3.17). To get eq. (3.18), we note that although the velocity is not constant, for a velocity changing at a *constant rate*, the displacement

will still be the *average* velocity times the time interval:

$$\text{displacement} = \text{average velocity} \times \text{time}$$
$$x - x_0 = \tfrac{1}{2}(v_0 + v)t$$

If we substitute $v_0 + at$ for v, we again arrive at eq. (3.18):

$$x - x_0 = \tfrac{1}{2}(v_0 + v_0 + at)t = v_0 t + \tfrac{1}{2}at^2$$

3.2.2 Vertical Free-Fall

Objects, as you may have noticed, tend to fall downward under the influence of gravity. By free-fall, we mean that nothing (such as air resistance) is opposing that fall. For small, heavy objects and reasonable velocities, ignoring air resistance is not too bad an approximation.

As we will see when we get to Newton's law of gravity, as long as you are near the Earth's surface, the downward acceleration g due to gravity is very nearly constant, at about $9.8\,\mathrm{m/s^2}$.[14] We can therefore apply the constant acceleration equations (3.17)-(3.21) to the case of free-fall. We could apply these relations exactly as we wrote them in the preceding section, but for the case of free-fall it is conventional to use a y axis for the vertical direction and to take up to be the positive y direction. This means replacing x by y and also setting $a = -g$ in eqq. (3.17)-(3.21).[15] Thus we have

$$v = v_0 - gt \tag{3.22a}$$
$$y - y_0 = v_0 t - \tfrac{1}{2}gt^2 \tag{3.22b}$$
$$= vt + \tfrac{1}{2}gt^2 \tag{3.22c}$$
$$= \frac{v + v_0}{2}\, t \tag{3.22d}$$
$$v^2 - v_0^2 = -2g(y - y_0) \tag{3.22e}$$

as you should be able to work out from eqq. (3.17)-(3.21) any time you need them.

Note that once we have chosen to make down the negative direction, displacements and velocities must follow this convention as well: a positive velocity is upward, an object that falls downward has a negative displacement, etc. Be careful not to muck up your signs; this is by far the most commonly committed error.

[14] The acceleration g due to gravity would be constant at least for points at sea level if the Earth were an isolated, nonrotating perfect fluid, but there are variations in both the magnitude and the direction of g due to the Earth's rotation and inhomogeneous geology. A freely distributable, 300-page scholarly treatise on the shape of the Earth and variations in g can be found at `http://samizdat.mines.edu/geodesy/geodesy.ps.gz`.

[15] The universal convention is that g stands for the *magnitude* of the acceleration, and we therefore have to put in any signs needed for the direction by hand.

3.3 Two-Dimensional Motion

As we saw in §3.1, in higher dimensions the relations for velocity and acceleration separate by components. In the case of two dimensions, we have, from eqq. (3.5) and (3.9),

$$a_x = \frac{dv_x}{dt} \qquad v_x = \frac{dx}{dt} \tag{3.23a}$$

$$a_y = \frac{dv_y}{dt} \qquad v_y = \frac{dy}{dt} \tag{3.23b}$$

Thus we have simply two separate sets of effectively one-dimensional equations, which can be formally solved by exactly the same methods as the one-dimensional case discussed in §3.2:

$$v_x - v_{0x} = \int_0^t a_x\,dt \qquad x - x_0 = \int_0^t v_x\,dt \tag{3.24a}$$

$$v_y - v_{0y} = \int_0^t a_y\,dt \qquad y - y_0 = \int_0^t v_y\,dt \tag{3.24b}$$

You should, however, note that although the relations for displacement, velocity, and acceleration separate by component, the motions in the x and y directions are not independent: they are coupled by their common parametrization by the time t.

You should also note that since in higher dimensions vector quantities have more than one component, directions can no longer be indicated simply by $\pm$ signs and magnitudes are no longer simply absolute values. Directions must now be specified in terms of angles, and magnitudes must be determined by the Pythagorean theorem. For the velocity vector $\mathbf{v}$, for example, we have

$$v_x = v\cos\theta \qquad v_y = v\sin\theta$$

$$v = \sqrt{v_x^2 + v_y^2} \qquad \tan\theta = \frac{v_y}{v_x}$$

and likewise for the acceleration or any other vector quantity. In particular bear in mind that it is the magnitude of the velocity, as given by the above Pythagorean relation, that tells how fast the body is moving.

3.3.1 Projectile Motion

By projectile motion, we mean the motion of bodies falling freely under the influence of gravity, but not restricted to the vertical: there may be horizontal as well as vertical motion. Once again we will for simplicity neglect the effects of air resistance.[16]

[16] For the curious, the case of projectile motion with air resistance is dealt with in Appendix A.

Projectile motion is a special case of two-dimensional motion. In the vertical direction there is a constant downward acceleration $a_y = -g$ due to gravity and in the horizontal direction there is no acceleration ($a_x = 0$). Thus eqq. (3.24a) become

$$v_x - v_{0x} = 0 \quad \Rightarrow \quad v_x = v_{0x} \tag{3.25a}$$

$$x - x_0 = \int_0^t v_x\,dt = \int_0^t v_{0x}\,dt = v_{0x}t \tag{3.25b}$$

where we have, in eq. (3.25b), substituted $v_x = v_{0x}$ from eq. (3.25a) and noted, in doing out the integration, that v_{0x} is a constant. Physically, eqq. (3.25) are telling us that in the horizontal direction we have simply motion at a constant speed — v_x is constant at whatever value v_{0x} it had initially — and that the displacement $x - x_0$ is therefore just rate times time.

Using $a_y = -g$ and the methods of §3.2.1 in eqq. (3.24b), we have

$$v_y - v_{0y} = \int_0^t -g\,dt = -gt \quad \Rightarrow \quad v_y = v_{0y} - gt \tag{3.26a}$$

$$y - y_0 = \int_0^t v_y\,dt = \int_0^t (v_{0y} - gt)\,dt = v_{0y}t - \tfrac{1}{2}gt^2 \tag{3.26b}$$

Comparing eqq. (3.26) with eqq. (3.22a) and (3.22b), we see that in the vertical direction we have the same relations that we had for vertical free-fall, even though for projectiles the motion is not purely vertical.

Projectile motion is thus just a superposition vertical free-fall with horizontal motion at a constant velocity. Using eqq. (3.25) or (3.26), we can calculate the horizontal and vertical motions independently of each other. But though we can *calculate* the horizontal and vertical motions independently, these two motions are in fact coupled by their common parametrization by the time t: the horizontal motion at constant velocity is occurring simultaneously with the vertical free-fall. In fact, we can combine the relations that eq. (3.25b) and (3.26b) give for the horizontal and vertical displacements $x - x_0$ and $y - y_0$ to eliminate t and obtain a result for the path $y = y(x)$ followed by the projectile: $x - x_0 = v_{0x}t$, when solved for t, yields $t = (x - x_0)/v_{0x}$. Substituting this into relation (3.26b) for $y - y_0$ then gives

$$\begin{aligned} y - y_0 &= v_{0y}t - \tfrac{1}{2}gt^2 \\ &= v_{0y}\,\frac{x - x_0}{v_{0x}} - \tfrac{1}{2}g\left(\frac{x - x_0}{v_{0x}}\right)^2 \end{aligned}$$

We could do some algebra to beautify this result, but we can already see that y is a quadratic function of x. The trajectory is therefore parabolic.

The highest point on the trajectory is called the apex. A common error is to think that the velocity vanishes at the apex. In fact, the projectile has the same *horizontal* velocity $v_x = v_{0x}$ at the apex that it has everywhere else

along the trajectory; what vanishes at the apex is the *vertical* component of the velocity: the projectile has ceased rising and is about to fall back down. You will deal with the apex in problem #46.

All of the information about a projectile's motion is in eqq. (3.25) and (3.26). We can, however, combine these relations to obtain one analogous to eq. (3.21):

$$\begin{aligned} v^2 - v_0^2 &= (v_x^2 + v_y^2) - (v_{0x}^2 + v_{0y}^2) \\ &= \Big(v_{0x}^2 + (v_{0y} - gt)^2\Big) - (v_{0x}^2 + v_{0y}^2) \\ &= v_{0x}^2 + v_{0y}^2 - 2v_{0y}gt + g^2t^2 - v_{0x}^2 - v_{0y}^2 \\ &= -2v_{0y}gt + g^2t^2 \\ &= -2g\Big(v_{0y}t - \tfrac{1}{2}gt^2\Big) \end{aligned}$$

The expression in parentheses is exactly that for $y - y_0$, so that we obtain

$$v^2 - v_0^2 = -2g(y - y_0) \tag{3.27}$$

Note that the right-hand side of this relation involves only the *vertical* displacement. Why this is so will be more clear when we cover conservation of energy.

While we're at it, you should also be comfortable deriving the projectile equations of motion in a purely vector formalism: the acceleration being g in the $-\hat{\mathbf{y}}$ direction, we have as our starting point

$$\mathbf{a} = -g\,\hat{\mathbf{y}}$$

Since this acceleration is constant, we can integrate it to obtain a result for the velocity $\mathbf{v}$:

$$\begin{aligned} \mathbf{a} = \frac{d\mathbf{v}}{dt} &= -g\,\hat{\mathbf{y}} \\ \int_{\mathbf{v}_0}^{\mathbf{v}} d\mathbf{v} &= \int_0^t dt\,(-g\,\hat{\mathbf{y}}) \\ \mathbf{v} - \mathbf{v}_0 &= -g\,\hat{\mathbf{y}}\,t\Big|_0^t = -gt\,\hat{\mathbf{y}} \\ \mathbf{v} &= \mathbf{v}_0 - gt\,\hat{\mathbf{y}} \end{aligned}$$

The x and y components of this result for $\mathbf{v}$ are simply the velocity relations (3.25a) and (3.26a), and it can be further integrated to obtain a result for the location $\mathbf{r}$ of the projectile:

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \mathbf{v}_0 - gt\,\hat{\mathbf{y}}$$

$$\int_{\mathbf{r}_0}^{\mathbf{r}} d\mathbf{r} = \int_0^t dt\,(\mathbf{v}_0 - gt\,\hat{\mathbf{y}})$$
$$\mathbf{r} - \mathbf{r}_0 = \mathbf{v}_0 t - \tfrac{1}{2}gt^2\,\hat{\mathbf{y}}$$

The x and y components of this result for $\mathbf{r} - \mathbf{r}_0$ are simply the relations (3.25b) and (3.26b) for the displacements $x - x_0$ and $y - y_0$.

3.3.2 Uniform Circular Motion

We will postpone consideration of what causes a body to move in a circular path until we cover dynamics; for now, we are concerned only to describe the motion quantitatively.

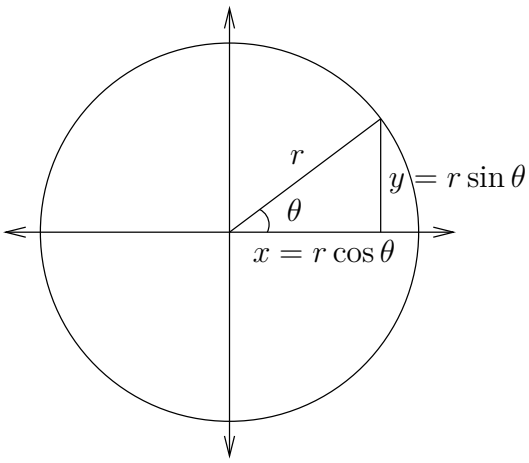


Figure 3.5: Parametrizing Circular Motion

For a body following the circular path shown in fig. (3.5), the radius r (in other words, the magnitude of its position vector $\mathbf{r}$) is constant; what changes is the angle θ at which the body is located. It is therefore natural to describe circular motion in terms of angular quantities. For linear motion, we described the location of the body by an x coordinate, corresponding to which were a velocity and an acceleration given by

$$v = \frac{dx}{dt} \qquad\qquad a = \frac{dv}{dt} = \frac{d^2x}{dt^2}$$

Since we are now describing the location of the body by a θ coordinate, we will, by analogy, define angular velocity ω and angular acceleration α by [17]

$$\omega = \frac{d\theta}{dt} \qquad\qquad \alpha = \frac{d\omega}{dt} = \frac{d^2\theta}{dt^2} \tag{3.28}$$

[17] This is ω, the Greek letter omega, not w; in handwriting, w has pointy bottoms, ω is rounded. A savage beating with the foam noodle awaits those who would confuse omega with "double-u."

Also, since angles are naturally measured in radians, the MKS units of ω are rad/sec.

The linear velocity v and the angular velocity ω both tell us how fast the body is moving; it is just that the angular velocity ω does so in terms of angle: ω is the angular rate at which the body is revolving around the circle. Similarly, the angular acceleration α tells us how rapidly the angular velocity of the body is changing — how quickly the body is speeding up or slowing down as it revolves.

The object of the game is now to get results for the vector velocity $\mathbf{v}$ and acceleration $\mathbf{a}$ of the body. We will do this by first finding an expression for the body's position vector $\mathbf{r}$ and then taking derivatives to get $\mathbf{v} = d\mathbf{r}/dt$ and $\mathbf{a} = d\mathbf{v}/dt$.

In this section we will restrict ourselves to the special case of uniform circular motion, which means that the rate ω at which the body is revolving is constant. Since $\omega = d\theta/dt$ is constant, we have

$$\int_{\theta_0}^{\theta} d\theta = \int_0^t \omega \, dt$$
$$\theta - \theta_0 = \omega t$$

Although the body could of course be at any angle θ_0 at time $t = 0$, we will for simplicity suppose that $\theta_0 = 0$.[18] Then $\theta = \omega t$ and, as you can see from fig. (3.5), we have

$$x = r\cos\theta = r\cos\omega t$$
$$y = r\sin\theta = r\sin\omega t$$

The position vector $\mathbf{r}$ of the body is therefore (recall eq. (3.1))

$$\begin{aligned}\mathbf{r} &= x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} \\ &= r\cos\theta\,\hat{\mathbf{x}} + r\sin\theta\,\hat{\mathbf{y}} \\ &= r\cos\omega t\,\hat{\mathbf{x}} + r\sin\omega t\,\hat{\mathbf{y}}\end{aligned} \tag{3.29}$$

Differentiating this, and remembering that r, ω, and the unit vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are constant, we have, by the chain rule,

$$\begin{aligned}\mathbf{v} &= \frac{d}{dt}(r\cos\omega t\,\hat{\mathbf{x}} + r\sin\omega t\,\hat{\mathbf{y}}) \\ &= r\,\frac{d(\cos\omega t)}{dt}\,\hat{\mathbf{x}} + r\,\frac{d(\sin\omega t)}{dt}\,\hat{\mathbf{y}} \\ &= r(-\omega\sin\omega t)\,\hat{\mathbf{x}} + r(\omega\cos\omega t)\,\hat{\mathbf{y}}\end{aligned}$$

(Although a radian is not really a physical unit, so that its MKS units could be written simply as $\sec^{-1}$.) Similarly, the MKS units of the angular acceleration α are $\mathrm{rad/sec^2}$ (or $\sec^{-2}$).

[18] This will not cause any loss of generality; we can always set up our coordinate system so that at time $t = 0$ the body is lined up with $\theta = 0$.

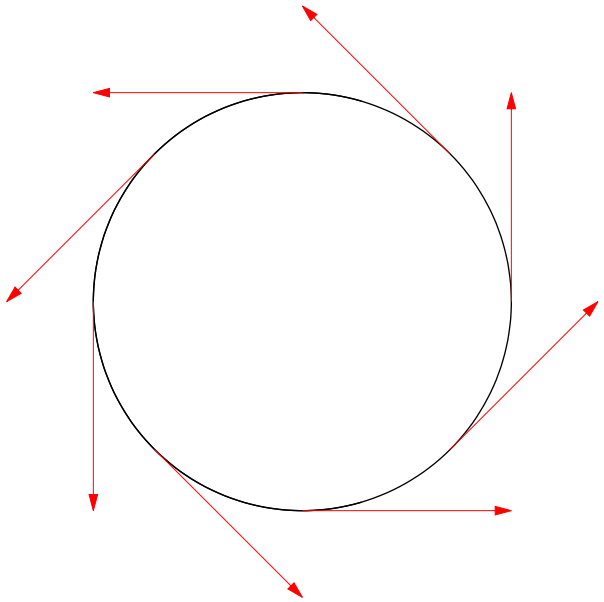


Figure 3.6: Direction of the Velocity

$$= \omega r(-\sin\omega t\,\hat{\mathbf{x}} + \cos\omega t\,\hat{\mathbf{y}}) \tag{3.30}$$

Now, we already know that the direction of the velocity $\mathbf{v}$ is, as always, tangent to the path followed by the body, as shown by the red arrows in fig. (3.6) for the case that the body is revolving in a counterclockwise circle. The speed (that is, the magnitude of the velocity) v will be

$$\begin{aligned} v &= \sqrt{v_x^2 + v_y^2} \\ &= \sqrt{(-\omega r\sin\omega t)^2 + (\omega r\cos\omega t)^2} \\ &= \omega r\sqrt{\sin^2\omega t + \cos^2\omega t} \\ &= \omega r \end{aligned} \tag{3.31}$$

Since ω and r are both constant, so is v: as we would expect for uniform circular motion, the speed of the body is constant.

For the acceleration we have

$$\begin{aligned} \mathbf{a} &= \frac{d\mathbf{v}}{dt} \\ &= \frac{d}{dt}[\omega r(-\sin\omega t\,\hat{\mathbf{x}} + \cos\omega t\,\hat{\mathbf{y}})] \\ &= \omega r\left[-\frac{d(\sin\omega t)}{dt}\,\hat{\mathbf{x}} + \frac{d(\cos\omega t)}{dt}\,\hat{\mathbf{y}}\right] \end{aligned}$$

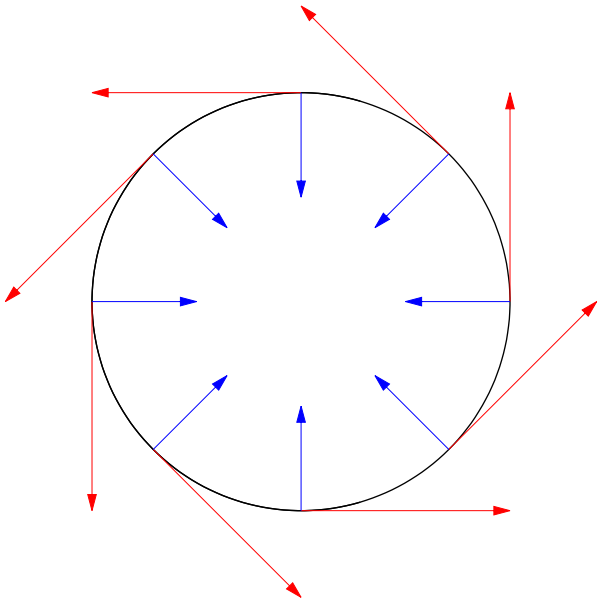


Figure 3.7: Direction of the Acceleration

$$\begin{aligned} &= \omega r[-(\omega\cos\omega t)\,\hat{\mathbf{x}} + (-\omega\sin\omega t)\,\hat{\mathbf{y}}] \\ &= -\omega^2(r\cos\omega t\,\hat{\mathbf{x}} + r\sin\omega t\,\hat{\mathbf{y}}) \\ &= -\omega^2\mathbf{r} \end{aligned} \tag{3.32}$$

The magnitude of the acceleration is thus $\omega^2 r$. Since $-\omega^2\mathbf{r}$ is in the direction opposite to $\mathbf{r}$, this means that the acceleration is directed in toward the center of the circle, as shown by the blue arrows in fig. (3.7). For this reason, it is called the centripetal acceleration (*centripetal* being Latin for *center-seeking*).

To summarize, by simply setting up an expression for the body's position vector and applying the definitions of the velocity and acceleration, we have obtained the following results for the case of uniform circular motion:

- The velocity of the body is tangent to the circle. The speed is $v = \omega r$.
- The acceleration of the body is in toward the center of the circle (centripetal acceleration). Its magnitude is $a = \omega^2 r$.

Note that uniform circular motion is an example of a case for which there is a nonzero acceleration even though the body's speed is constant. As you can see from fig. (3.6), the direction of the velocity is continually changing as the body revolves around the circle; its acceleration is due entirely to this change in direction. And if you think about it a bit, it should even seem

plausible that the velocity vector is continually being "bent" in toward the center of the circle, and hence that the acceleration is in that direction.

In addition to the above results for velocity and acceleration, there are several other quantities conventionally used to describe and analyze circular motion:

- Each complete circle made by the body can also be termed a cycle.
- The period T of the motion is the time to complete one full cycle.
- The frequency f of the motion is the number of cycles per unit time.
  - Since the period T is time per cycle and f is cycles per unit time, the frequency and the period are reciprocals of each other:

$$f = \frac{1}{T} \tag{3.33}$$

  - The MKS units of frequency are cycles per second (cps) or Hertz (Hz), which is just another name for the same thing: 1 cps = 1 Hz.
  - The frequency f, like the angular velocity ω, indicates how rapidly the body is revolving. The difference between f and ω is that f measures this rate in terms of cycles while ω measures it in terms of angle (radians). Because there are 2π radians in a full cycle, there is a simple relation between f and ω:

$$\frac{\text{radians}}{\text{sec}} = \frac{2\pi \text{ rad}}{\text{cycle}} \times \frac{\text{cycles}}{\text{sec}}$$
$$\omega = 2\pi f \tag{3.34}$$

    Just to confuse matters, because of the close relationship between the frequency f and the angular velocity ω, ω is often also called the angular frequency. You must therefore be very careful to distinguish between *angular frequency*, which means ω, and just plain old *frequency*, which means f.

We could also have arrived at the result $v = \omega r$, at least for the case of uniform circular motion, simply by reasoning it out: since the body is revolving at a constant rate, the speed will be constant, which means that there is no difference between the instantaneous and average speed. If we average over a full cycle,

$$\text{speed} = \frac{\text{circumference}}{\text{period}}$$
$$v = \frac{2\pi r}{T} = 2\pi r f = (2\pi f) r = \omega r$$

Finally, combining $v = \omega r$ with $a = \omega^2 r$, we have

$$a = \omega^2 r = \frac{(\omega r)^2}{r} = \frac{v^2}{r}$$

$a = \omega^2 r$ and $a = v^2/r$ are, of course, completely equivalent, but sometimes one is more convenient for the purposes of calculation than the other.

3.3.3 Centripetal Acceleration, Newton-Style

Although Newton invented calculus, his derivation of the result for centripetal acceleration in his great work *Philosophiæ Naturalis Principia Mathematica* was geometrical: he considered, as an approximation to circular motion, a point-like object tracing out a polygon of n sides as it bounced around the inside of a circle, reasoning that in the limit that $n \to \infty$ this polygonal motion would reproduce circular motion. If we think of the circular perimeter as a "brick wall," on each bounce the component of the object's velocity parallel to the wall (that is, the component tangent to the circle) will be unaffected, while the component perpendicular to the wall (the component along the radius of the circle) will reverse direction.

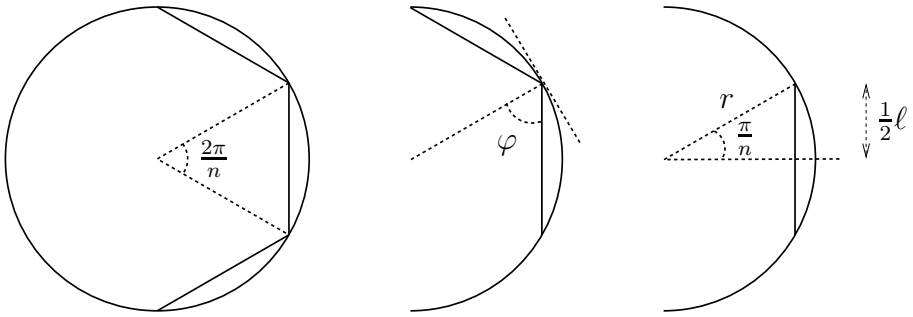


Figure 3.8: Newton's Polygonal Path

As you can see from the left side of fig. (3.8), each side of the polygon will subtend a central angle of $2\pi/n$, so that each of the other angles in the isosceles triangle formed with a side of the polygon will be

$$\varphi = \frac{1}{2}\left(\pi - \frac{2\pi}{n}\right) = \frac{\pi}{2} - \frac{\pi}{n}$$

And while we are at it, the angle in the triangle shown on the right side of fig. (3.8) will be half of the central angle, and since the vertical side of that right triangle is half of the length ℓ of a side of the polygon, we also have

$$\tfrac{1}{2}\ell = r \sin \frac{\pi}{n}$$

which gives us $\ell = 2r \sin \frac{\pi}{n}$ for the length of each side.

If the object is traveling at speed v up the side of the polygon we have drawn vertically, then, as can sort of be seen from the middle of fig. (3.8), the component of its velocity along the radial direction, being the side adjacent to the angle φ, will be

$$v_{radial} = v \cos \varphi = v \cos\left(\frac{\pi}{2} - \frac{\pi}{n}\right) = v \sin \frac{\pi}{n}$$

On each bounce, it is this radial component of the object's velocity that is reversed: it will go from being $v \sin \frac{\pi}{n}$ radially outward to $v \sin \frac{\pi}{n}$ radially inward. If we consider radially outward to be our positive direction, the change in the object's velocity as a result of the bounce is therefore

$$\Delta v_{radial} = \left(-v \sin \frac{\pi}{n}\right) - \left(+v \sin \frac{\pi}{n}\right) = -2v \sin \frac{\pi}{n}$$

that is, the change in the object's velocity as a vector is $2v \sin \frac{\pi}{n}$ directed radially inward. Since there is one bounce for each vertex of the polygon and therefore for each side of the polygon, the time interval Δt during which this change in velocity occurs is just the time it takes the object to traverse the length ℓ of a side of the polygon at speed v:

$$\Delta t = \frac{\ell}{v}$$

and hence, since $\ell = 2r \sin \frac{\pi}{n}$,

$$\Delta t = \frac{2r \sin \frac{\pi}{n}}{v}$$

For the object's average acceleration during a bounce we thus arrive at

$$a_{avg} = \frac{\Delta v}{\Delta t} = \frac{2v \sin \frac{\pi}{n}}{\frac{2r \sin \frac{\pi}{n}}{v}} = \frac{v^2}{r}$$

which as a vector is, like Δv, radially inward toward the center of the circle. While our result for this average acceleration is curiously independent of the number of sides n of the polygon, it is in the limit $n \to \infty$ that the object's polygonal motion will become circular and our result for its average acceleration will become the instantaneous centripetal acceleration of an object in uniform circular motion.

3.3.4 Nonuniform Circular Motion

We will now allow the rate at which the body is revolving around the circle to vary. This means that the angular velocity ω will no longer be constant, so that θ is no longer simply ωt. We must therefore leave our expression (3.29) for the position vector in the more general form

$$\mathbf{r} = r\cos\theta\,\hat{\mathbf{x}} + r\sin\theta\,\hat{\mathbf{y}} \tag{3.35}$$

To obtain a result for the velocity $\mathbf{v}$, we once again take the derivative of the position vector:

$$\begin{aligned}
\mathbf{v} &= \frac{d\mathbf{r}}{dt} \\
&= \frac{d}{dt}(r\cos\theta\,\hat{\mathbf{x}} + r\sin\theta\,\hat{\mathbf{y}}) \\
&= r\,\frac{d(\cos\theta)}{dt}\,\hat{\mathbf{x}} + r\,\frac{d(\sin\theta)}{dt}\,\hat{\mathbf{y}} \\
&= r\left(-\frac{d\theta}{dt}\sin\theta\right)\,\hat{\mathbf{x}} + r\left(\frac{d\theta}{dt}\cos\theta\right)\,\hat{\mathbf{y}} \\
&= r(-\omega\sin\theta)\,\hat{\mathbf{x}} + r(\omega\cos\theta)\,\hat{\mathbf{y}} \\
&= \omega r(-\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}})
\end{aligned}$$

Except that we have not used ωt for θ, this is exactly the same result that we obtained in the case of uniform circular motion (eq. (3.30)), so that, as before, the direction of the velocity is tangent to the circle and the speed is given by $v = \omega r$. The difference this time is that ω and v are not constant — which has consequences when we now take the derivative of the velocity to get the acceleration:

$$\begin{aligned}
\mathbf{a} &= \frac{d\mathbf{v}}{dt} \\
&= \frac{d}{dt}\Big[\omega r(-\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}})\Big] \\
&= \omega r\left[-\frac{d(\sin\theta)}{dt}\,\hat{\mathbf{x}} + \frac{d(\cos\theta)}{dt}\,\hat{\mathbf{y}}\right] + \frac{d\omega}{dt}\,r(-\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}}) \\
&= \omega r\left[-\left(\frac{d\theta}{dt}\cos\theta\right)\,\hat{\mathbf{x}} + \left(-\frac{d\theta}{dt}\sin\theta\right)\,\hat{\mathbf{y}}\right] + \alpha r(-\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}}) \\
&= \omega r\Big[-(\omega\cos\theta)\,\hat{\mathbf{x}} + (-\omega\sin\theta)\,\hat{\mathbf{y}}\Big] + \alpha r(-\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}}) \\
&= -\omega^2(r\cos\theta\,\hat{\mathbf{x}} + r\sin\theta\,\hat{\mathbf{y}}) + \alpha r(-\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}}) \\
&= -\omega^2\mathbf{r} + \alpha r(-\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}})
\end{aligned} \tag{3.36}$$

The first term on the right-hand side is the same result that we obtained in the case of uniform circular motion: a centripetal acceleration $\omega^2 r$ directed in toward the center of the circle.

The new feature is the contribution from the terms with the angular acceleration α. The expression in parentheses, $-\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}}$, turns out to be a unit vector tangent to the circle in the counterclockwise direction and is conventionally denoted by $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = -\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}}$$

To verify that $\hat{\boldsymbol{\theta}}$ is tangential, we note that if we dot it with the radial vector $\mathbf{r}$, we have (see p.78)

$$\begin{aligned}
\mathbf{r}\cdot\hat{\boldsymbol{\theta}} &= (r\cos\theta\,\hat{\mathbf{x}} + r\sin\theta\,\hat{\mathbf{y}})\cdot(-\sin\theta\,\hat{\mathbf{x}} + \cos\theta\,\hat{\mathbf{y}}) \\
&= (r\cos\theta)(-\sin\theta)\,\hat{\mathbf{x}}\cdot\hat{\mathbf{x}} \\
&\qquad + \Big[(r\cos\theta)(\cos\theta) + (r\sin\theta)(-\sin\theta)\Big]\,\hat{\mathbf{x}}\cdot\hat{\mathbf{y}} \\
&\qquad + (r\sin\theta)(\cos\theta)\,\hat{\mathbf{y}}\cdot\hat{\mathbf{y}} \\
&= (r\cos\theta)(-\sin\theta)(1) \\
&\qquad + \Big[(r\cos\theta)(\cos\theta) + (r\sin\theta)(-\sin\theta)\Big](0) \\
&\qquad + (r\sin\theta)(\cos\theta)(1) \\
&= 0
\end{aligned}$$

This means that radial vector $\mathbf{r}$ and $\hat{\boldsymbol{\theta}}$ are perpendicular — so $\hat{\boldsymbol{\theta}}$ must be tangential. To verify that $\hat{\boldsymbol{\theta}}$ is tangential in the counterclockwise direction, we can simply evaluate it at $\theta = 0$:

$$\hat{\boldsymbol{\theta}}\Big|_{\theta=0} = (-\sin 0\,\hat{\mathbf{x}} + \cos 0\,\hat{\mathbf{y}}) = \hat{\mathbf{y}}$$

That is, when $\theta = 0$, $\hat{\boldsymbol{\theta}}$ is in the positive y direction, which is indeed counterclockwise. Finally, to verify that $\hat{\boldsymbol{\theta}}$ is in fact a unit vector, we check that it has unit magnitude:

$$\left|\hat{\boldsymbol{\theta}}\right| = \sqrt{\left[\hat{\boldsymbol{\theta}}\right]_x^2 + \left[\hat{\boldsymbol{\theta}}\right]_y^2} = \sqrt{(-\sin\theta)^2 + (\cos\theta)^2} = 1$$

The new contribution in eq. (3.36) may therefore be written as

$$\mathbf{a}_{tan} = \alpha r\hat{\boldsymbol{\theta}}$$

— that is, an acceleration of magnitude αr in the counterclockwise tangential direction.[19] Just as the angular acceleration α indicates how the body's

[19] Or of course clockwise if α is negative.

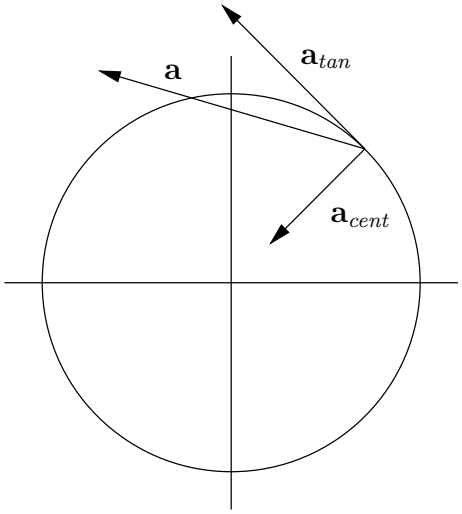


Figure 3.9: The Centripetal and Tangential Contributions to the Acceleration

angular rate of revolution ω is changing, the tangential acceleration $a_{tan} = \alpha r$ indicates to how the body's tangential velocity v is changing. That is, $a_{tan} = \alpha r$ corresponds to the body's speeding up or slowing down as it revolves around the circle.

This dichotomy in the acceleration, illustrated in fig. (3.9), is an example of the general result worked out at the end of §3.1: a body's acceleration consists of a component parallel to the velocity that corresponds to a change in the magnitude of the body's velocity (that is, to speeding up or slowing down) and a component perpendicular to the velocity that corresponds to a change in the direction of the body's velocity. In the case of circular motion, the tangential acceleration corresponds to the body's change in speed and the centripetal acceleration to the continual change in the body's direction of motion as it veers around in a circular path.

To summarize, in the case of nonuniform circular motion, we have the following results:

- The velocity of the body is tangent to the circle. The speed, whether constant or changing, is $v = \omega r$.
- There are two contributions to the acceleration of the body, one radial and the other tangential:

$$\mathbf{a} = \mathbf{a}_{cent} + \mathbf{a}_{tan}$$

  - The radial part of the acceleration is in toward the center of the circle (centripetal acceleration). Its magnitude, whether ω constant or changing, is $a_{cent} = \omega^2 r$.
  - The tangential part of the acceleration, $a_{tan} = \alpha r$, corresponds to the body's speeding up or slowing down as it revolves.

Finally, note that there is a simple hierarchy of relations among tangential and angular quantities in terms of the radius r of the circle: the arc length s subtended by angle θ in a circle of radius r is $s = \theta r$ and corresponds to the distance the body has traveled along the perimeter of the circle. We therefore expect the rate of change of this distance, ds/dt, to correspond to the speed of the body, and in fact, since r is constant,

$$\frac{ds}{dt} = \frac{d(\theta r)}{dt} = \frac{d\theta}{dt}\, r = \omega r = v$$

If we are concerned only with the part of the acceleration due to changes in speed, we also expect that we can get this by taking the derivative of the speed $v = \omega r$, and in fact

$$\frac{dv}{dt} = \frac{d(\omega r)}{dt} = \frac{d\omega}{dt}\, r = \alpha r = a_{tan}$$

Thus we have the simple hierarchy

$$s = \theta r \qquad\qquad v = \omega r \qquad\qquad a_{tan} = \alpha r$$

3.3.5 General Motion in Polar Coordinates

A general vector, say $\mathbf{H}$, has in polar coordinates (r, θ) both $\hat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$ components:

$$\mathbf{H} = h_r \hat{\mathbf{r}} + h_\theta \widehat{\boldsymbol{\theta}}$$

One complication when working in polar coordinates is that the unit vectors $\hat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$, unlike $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, are *not* constant: while of course $\hat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$ both always have unit magnitude, their directions depend on where you are. As you can, by considering the projections of $\hat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$ onto $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ and vice versa, see from fig. (3.10) with a bit of imagination and some luck,[20]

$$\hat{\mathbf{r}} = \cos\theta\, \hat{\mathbf{x}} + \sin\theta\, \hat{\mathbf{y}} \tag{3.37a}$$

$$\widehat{\boldsymbol{\theta}} = -\sin\theta\, \hat{\mathbf{x}} + \cos\theta\, \hat{\mathbf{y}} \tag{3.37b}$$

and

[20] An easy way to reproduce these relations from scratch is to remember that all of the coefficients must be either $\pm\sin\theta$ or $\pm\cos\theta$. To sort out the signs and the $\sin\theta$'s versus the $\cos\theta$'s, note first that because of the right angle between the two contributions being combined, one term will always involve $\sin\theta$ and the other $\cos\theta$, then think of an extreme case. At $\theta = 0$, for example, we should have $\widehat{\boldsymbol{\theta}} = +\hat{\mathbf{y}}$, which tells us that the coefficient of $\hat{\mathbf{y}}$ in the expression for $\widehat{\boldsymbol{\theta}}$ should be $+\cos\theta$ and that therefore the coefficient of $\hat{\mathbf{x}}$ must be $\pm\sin\theta$. To see that we want $-\sin\theta\,\hat{\mathbf{x}}$ rather than $+\sin\theta\,\hat{\mathbf{x}}$, we can think of fig. (3.10) and see that $\widehat{\boldsymbol{\theta}}$ is tilted in the negative x direction, or we can look at another extreme case: when $\theta = \frac{\pi}{2}$, we should have $\widehat{\boldsymbol{\theta}} = -\hat{\mathbf{x}}$.

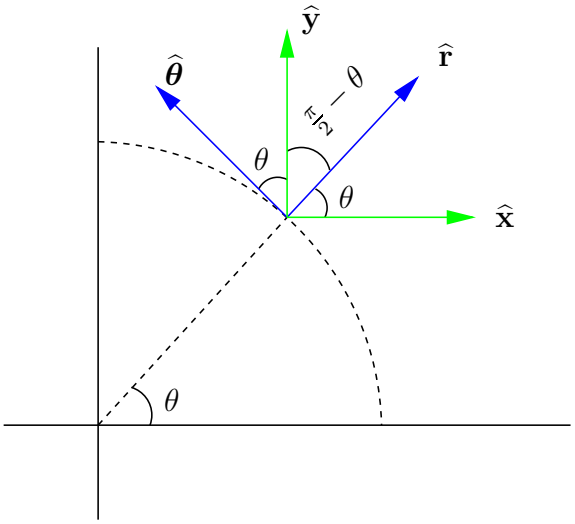


Figure 3.10: Relations Among Unit Vectors

$$\widehat{\mathbf{x}} = \cos\theta\,\widehat{\mathbf{r}} - \sin\theta\,\widehat{\boldsymbol{\theta}} \tag{3.38a}$$

$$\widehat{\mathbf{y}} = \sin\theta\,\widehat{\mathbf{r}} + \cos\theta\,\widehat{\boldsymbol{\theta}} \tag{3.38b}$$

In eqq. (3.37), although the $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}$ are constant, the value of θ, and thus the directions of $\widehat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$, depend on where you are.

In the context of kinematics, we will be taking time derivatives to obtain results for velocity and acceleration. Since θ is in general a function of time (that is, in general θ will change as the body moves around), and since $\widehat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$ depend on θ, the time derivatives of $\widehat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$, unlike those of $\widehat{\mathbf{x}}$ and $\widehat{\mathbf{y}}$, do not vanish: thus, for the above vector $\mathbf{H}$,

$$\frac{d\mathbf{H}}{dt} = \frac{d}{dt}(h_r\widehat{\mathbf{r}} + h_\theta\widehat{\boldsymbol{\theta}}) = \frac{dh_r}{dt}\,\widehat{\mathbf{r}} + h_r\frac{d\,\widehat{\mathbf{r}}}{dt} + \frac{dh_\theta}{dt}\,\widehat{\boldsymbol{\theta}} + h_\theta\frac{d\widehat{\boldsymbol{\theta}}}{dt}$$

We are therefore going to want general results for the time derivatives of $\widehat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$, so that we don't have to go through the bother of working them out from scratch all the time. Using eqq. (3.37), we have

$$\begin{aligned}\frac{d\,\widehat{\mathbf{r}}}{dt} &= \frac{d}{dt}(\cos\theta\,\widehat{\mathbf{x}} + \sin\theta\,\widehat{\mathbf{y}}) \\ &= \frac{d(\cos\theta)}{dt}\,\widehat{\mathbf{x}} + \frac{d(\sin\theta)}{dt}\,\widehat{\mathbf{y}} \\ &= -\sin\theta\,\frac{d\theta}{dt}\,\widehat{\mathbf{x}} + \cos\theta\,\frac{d\theta}{dt}\,\widehat{\mathbf{y}}\end{aligned} \qquad \begin{aligned}\frac{d\widehat{\boldsymbol{\theta}}}{dt} &= \frac{d}{dt}(-\sin\theta\,\widehat{\mathbf{x}} + \cos\theta\,\widehat{\mathbf{y}}) \\ &= -\frac{d(\sin\theta)}{dt}\,\widehat{\mathbf{x}} + \frac{d(\cos\theta)}{dt}\,\widehat{\mathbf{y}} \\ &= -\cos\theta\,\frac{d\theta}{dt}\,\widehat{\mathbf{x}} - \sin\theta\,\frac{d\theta}{dt}\,\widehat{\mathbf{y}}\end{aligned}$$

Since $d\theta/dt = \omega$, these relations reduce to

$$\frac{d\,\widehat{\mathbf{r}}}{dt} = -\sin\theta\,\omega\,\widehat{\mathbf{x}} + \cos\theta\,\omega\,\widehat{\mathbf{y}} \qquad \frac{d\,\widehat{\boldsymbol{\theta}}}{dt} = -\cos\theta\,\omega\,\widehat{\mathbf{x}} - \sin\theta\,\omega\,\widehat{\mathbf{y}}$$
$$= \omega(-\sin\theta\,\widehat{\mathbf{x}} + \cos\theta\,\widehat{\mathbf{y}}) \qquad = -\omega(\cos\theta\,\widehat{\mathbf{x}} - \sin\theta\,\widehat{\mathbf{y}})$$

From eqq. (3.38), we can see that these results, re-expressed in polar form, are simply [21]

$$\frac{d\,\widehat{\mathbf{r}}}{dt} = \omega\,\widehat{\boldsymbol{\theta}} \qquad \frac{d\,\widehat{\boldsymbol{\theta}}}{dt} = -\omega\,\widehat{\mathbf{r}} \tag{3.39}$$

We are now in a position to work out general results for the velocity and acceleration of a body in polar coordinates. The position vector $\widehat{\mathbf{r}}$ is purely radial, so it has no tangential ($\widehat{\boldsymbol{\theta}}$) component, and in polar coordinates the expression for it is simply [22]

$$\mathbf{r} = r\,\widehat{\mathbf{r}} \tag{3.40}$$

corresponding to going a distance r in the outward radial direction $\widehat{\mathbf{r}}$ to get to the location specified by $\mathbf{r}$. The velocity is therefore

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{d(r\,\widehat{\mathbf{r}})}{dt} = \frac{dr}{dt}\,\widehat{\mathbf{r}} + r\,\frac{d\,\widehat{\mathbf{r}}}{dt} = \frac{dr}{dt}\,\widehat{\mathbf{r}} + \omega r\,\widehat{\boldsymbol{\theta}}$$

Denoting a time derivative by placing a dot over a quantity (a notation that actually goes all the way back to Newton), our result for a general velocity in polar coordinates is

$$\mathbf{v} = \dot{r}\,\widehat{\mathbf{r}} + \omega r\,\widehat{\boldsymbol{\theta}} \tag{3.41}$$

The $\omega r\,\widehat{\boldsymbol{\theta}}$ term is just the tangential velocity we had arrived at in the case of circular motion. The new term $\dot{r}\,\widehat{\mathbf{r}}$ corresponds to radial motion, which is of course not present in purely circular motion since it involves a change of radius. This general result for velocity may also be written

$$\mathbf{v} = v_r\,\widehat{\mathbf{r}} + v_\theta\,\widehat{\boldsymbol{\theta}}$$

[21] We might have expected from the start that $d\widehat{\mathbf{r}} \propto \widehat{\boldsymbol{\theta}}$ and $d\widehat{\boldsymbol{\theta}} \propto \widehat{\mathbf{r}}$: although $\widehat{\mathbf{r}}$ and $\widehat{\boldsymbol{\theta}}$ are not constant, because they must remain unit vectors they can change only their directions and not their magnitude. The change in each must therefore be perpendicular to itself: in the $\pm\widehat{\boldsymbol{\theta}}$ direction for $\widehat{\mathbf{r}}$ and in the $\pm\widehat{\mathbf{r}}$ direction for $\widehat{\boldsymbol{\theta}}$.

[22] We could of course also construct this from the Cartesian form of $\mathbf{r}$ and eqq. (3.38):

$$\begin{aligned}\mathbf{r} &= x\,\widehat{\mathbf{x}} + y\,\widehat{\mathbf{y}}\\ &= r\cos\theta(\cos\theta\,\widehat{\mathbf{r}} - \sin\theta\,\widehat{\boldsymbol{\theta}}) + r\sin\theta(\sin\theta\,\widehat{\mathbf{r}} + \cos\theta\,\widehat{\boldsymbol{\theta}})\\ &= r(\cos^2\theta + \sin^2\theta)\widehat{\mathbf{r}}\\ &= r\,\widehat{\mathbf{r}}\end{aligned}$$

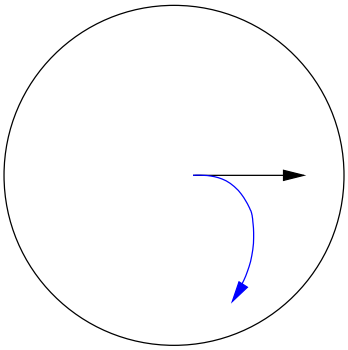


Figure 3.11: A Coriolis Acceleration. Sort of.

with

$$v_r = \dot{r} \qquad\qquad v_\theta = \omega r$$

To obtain a general result for acceleration in polar coordinates, we just take one more time derivative:

$$\begin{aligned}
\mathbf{a} &= \frac{d\mathbf{v}}{dt} \\
&= \frac{d}{dt}(\dot{r}\,\hat{\mathbf{r}} + \omega r\,\hat{\boldsymbol{\theta}}) \\
&= \frac{d\dot{r}}{dt}\,\hat{\mathbf{r}} + \dot{r}\,\frac{d\hat{\mathbf{r}}}{dt} + \frac{d\omega}{dt}\,r\hat{\boldsymbol{\theta}} + \omega\,\frac{dr}{dt}\,\hat{\boldsymbol{\theta}} + \omega r\,\frac{d\hat{\boldsymbol{\theta}}}{dt}
\end{aligned}$$

or, if we note that $d\omega/dt = \alpha$,

$$= \ddot{r}\,\hat{\mathbf{r}} + \dot{r}(\omega\,\hat{\boldsymbol{\theta}}) + \alpha r\hat{\boldsymbol{\theta}} + \omega\dot{r}\hat{\boldsymbol{\theta}} + \omega r(-\omega\,\hat{\mathbf{r}})$$

Combining like terms and beautifying, we arrive at

$$\mathbf{a} = \left(\ddot{r} - \omega^2 r\right)\hat{\mathbf{r}} + (\alpha r + 2\omega\dot{r})\,\hat{\boldsymbol{\theta}} \tag{3.42}$$

The $\omega^2 r\,\hat{\mathbf{r}}$ and $\alpha r\,\hat{\boldsymbol{\theta}}$ terms are just the centripetal and tangential accelerations we had arrived at in the case of circular motion. One of the new terms, the $\ddot{r}\,\hat{\mathbf{r}}$, corresponds to a simple radial acceleration — a speeding up or slowing down of the radial velocity. The other new term, the $2\omega\dot{r}\,\hat{\boldsymbol{\theta}}$, is the Coriolis acceleration and corresponds to a change in the body's radius of motion (by virtue of the $\dot{r}$) while the body is also moving tangentially (by virtue of the ω). Fig. (3.11) shows the Coriolis effect for a roach on a Lazy Susan that is rotating clockwise. We will suppose that initially the roach shares the angular velocity ω of the Lazy Susan and therefore has, at its radius r from the center, tangential velocity ωr. If the roach then tries to walk straight

radially outward (radial motion corresponding to $\dot{r}$), it will be moving out to larger r, but without any tangential acceleration to give a commensurate increase in its tangential velocity ωr. Relative to the Lazy Susan, the roach will therefore "fall behind" as it moves outward, following the spiral path shown in the figure.

Eq. (3.42) may also be written

$$\mathbf{a} = a_r\,\hat{\mathbf{r}} + a_\theta\,\hat{\boldsymbol{\theta}}$$

with

$$a_r = \ddot{r} - \omega^2 r \qquad\qquad a_\theta = \alpha r + 2\omega\dot{r}$$

3.3.6 Two-Dimensional Relative Velocities

As with relative velocities in one dimension, it is better to use intuition and common sense than to try to rely on a general formula. But you do need to be mindful that in two or more dimensions you must combine velocities as *vectors*.

If, for example, a plane is flying (in the sense of heading or "aiming") east at velocity $\mathbf{v}_p$ when there is a wind to the south at velocity $\mathbf{v}_w$, the plane will be carried along by the wind, so that relative to the ground its velocity will be $\mathbf{v}_g = \mathbf{v}_p + \mathbf{v}_w$, as shown in fig. (3.12). Since the velocities we are adding in this particular example are at right angles,

$$v_g = \sqrt{v_p^2 + v_w^2}$$

yada, yada, yada. (Although in more general cases the velocities need of course not be at right angles, and combining them would require doing a vector addition by the usual methods.)

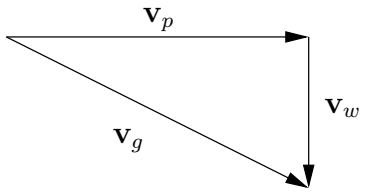


Figure 3.12: Combining Two-Dimensional Relative Velocities

3.4 Problems

1. At bath time, you escape and dash outside in your birthday suit, shrieking with joy as you run in circles of 60 ft diameter.[23]

   (a) If you complete one such circle in 20 sec,

      i. What is your average speed over this 20 sec interval?

      ii. What is your average velocity over this 20 sec interval?

   (b) If you complete a semicircle in 10 sec,

      i. What is your average speed over this 10 sec interval?

      ii. What is the magnitude of your average velocity over this 10 sec interval?

2. In the morning, a day student makes the 8 mi trip to school in 16 min. In the evening, that same student makes the return trip home at 120 mph.

   (a) At what average speed did the student make the morning trip?

   (b) What can you say about the instantaneous speed at each point along the way to school?

   (c) How long does the return trip take?

   (d) What is the average speed for the round trip?

3. At 8:08 one morning, you dash from your dorm to your first period class, covering 240 yd at 16 ft/sec, only to discover that the class has been canceled. You then cover the 120 yd to the dining hall at 4.0 ft/sec, kicking yourself all the way.

   (a) What time is it when you get to the classroom?

   (b) What is your average speed overall?

   (c) If instead you had run half the total of 360 yd at 16 ft/sec and walked the other half at 4.0 ft/sec, what would your overall average speed be?

   (d) If instead you had run half the *time* at 16 ft/sec and walked the other half at 4.0 ft/sec, what would your overall average speed be?

   (e) Why do your answers to #3c and #3d differ, and in particular why is the answer to #3c smaller than the answer to #3d?

[23] In your younger days, that is; for the purposes of the problem, we will assume that you are not still doing this sort of thing.

4. (a) If you cover half of a distance ℓ at speed v_1 and the other half at speed $v_2 \geq v_1$,
   i. What is your average speed?
   ii. How long does it take you to cover that total distance?
   iii. Is your average speed closer to v_1 or v_2? Why? See the footnote if you need a hint.[24]
   iv. Show that your average speed behaves as you would expect when $v_1 = v_2$.
   v. How does your average speed behave when v_2 is much larger than v_1?
   vi. What if v_1 is zero?

   (b) If instead you spend an equal time (that is, half the total time) at speeds v_1 and v_2 as you cover the same total distance ℓ,
   i. What is your average speed?
   ii. How long does it take you to cover that total distance?
   iii. Is this speed closer to v_1 or v_2? Why?
   iv. Show that your average speed behaves as you would expect when $v_1 = v_2$.
   v. How does your average speed behave when v_2 is much larger than v_1?
   vi. What if v_1 is zero?

5. A school bus averages 90 mph on its morning trip. The driver wants to make the afternoon trip fast enough to make his overall average speed for the day come out to 180 mph. Why isn't this possible? See the footnote if you need a hint.[25]

[24] One way to see whether your average speed v_{avg} is closer to v_1 or v_2 is to look at

$$\left| \frac{v_{avg} - v_1}{v_{avg} - v_2} \right|$$

[25] Set up the naive relations for the times and distances for each one-way trip and for the round trip.

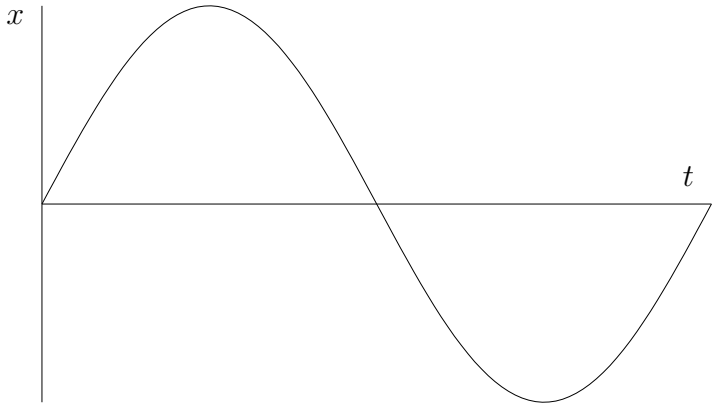


Figure 3.13: Problem 6

6. Fig. (3.13) shows a plot of the location x of a body along an x axis versus the time t. Make a sketch of the corresponding velocity v and acceleration a as functions of time. Describe the motion. That is, what would the body's motion actually look like?

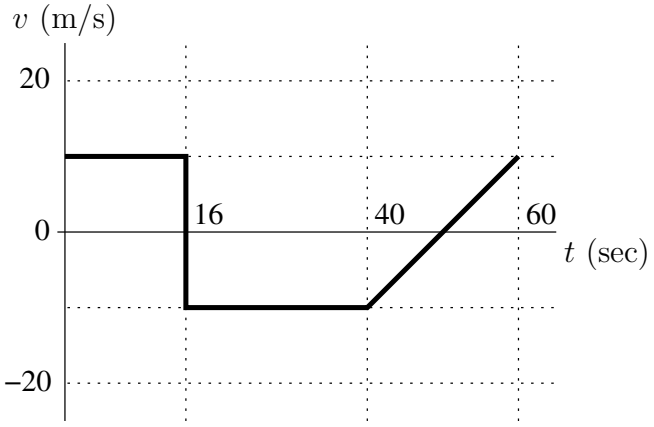


Figure 3.14: Problem 7

7. Fig. (3.14) shows the plot of velocity v versus time t for a cat being pursued with a pitchfork.

   (a) What is the cat's total displacement?

   (b) Describe the motion fully. That is, describe the cat's location, velocity, and acceleration, and how the cat's location, velocity, and acceleration are changing, at all times — whether the cat is moving forward or backward, how fast, whether speeding up or slowing down, that sort of thing.

   (c) What is the cat's overall average *speed*?

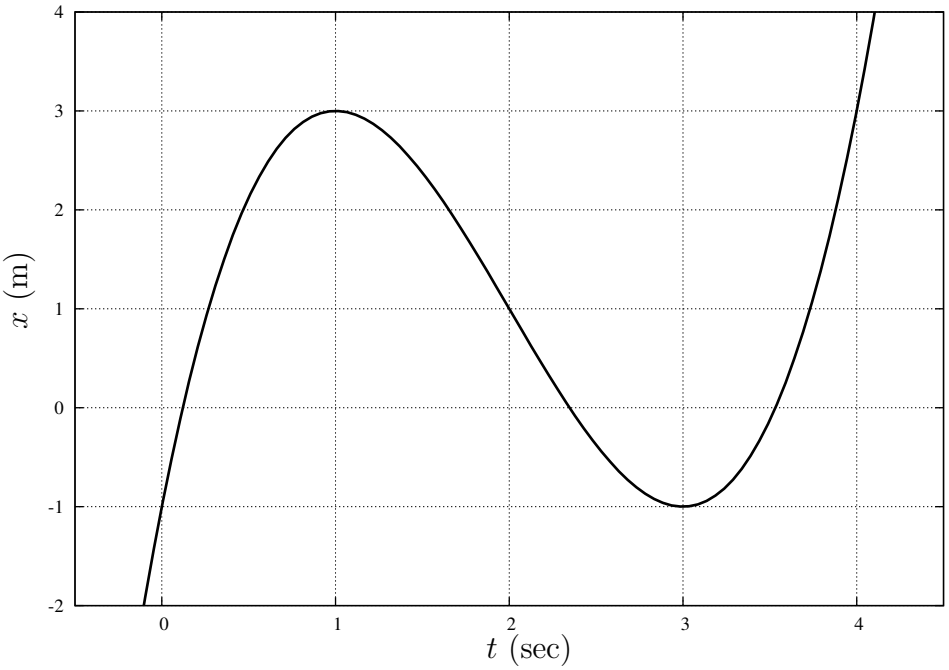


Figure 3.15: Problem 8

8. The position of some stupid object moving along an x axis is given by

$$x = t^3 - 6t^2 + 9t - 1$$

where x is in meters and t in seconds, as shown in fig. (3.15).

(a) What are the physical dimensions of the coefficients 1, -6, 9, and -1 in $x = t^3 - 6t^2 + 9t - 1$?[26]

(b) Describe the motion fully. That is, describe the object's location, velocity, and acceleration, and how the object's location, velocity, and acceleration are changing — whether it is moving forward or backward, how fast, whether speeding up or slowing down, etc. See the footnote if you want a hint that will leave you feeling like you've been gypped.[27]

[26]Such formulæ, with dimensionful numerical coefficients, are an abomination. They are legitimately used in pedagogical contexts in order to keep the algebra simple enough not to distract you from the physics involved, but be aware that coefficients with physical dimensions should properly be denoted by algebraic symbols because the numerical values of such coefficients depend on an arbitrary choice of units. The only numerical values that belong in physical formulæ are pure numbers like π and $\sqrt{2}$. (Engineers do occasionally use dimensionful numerical coefficients, but only for convenience in applying relations repeatedly to very specific contexts.)

[27]For the acceleration, think in terms of the inflection of the curve.

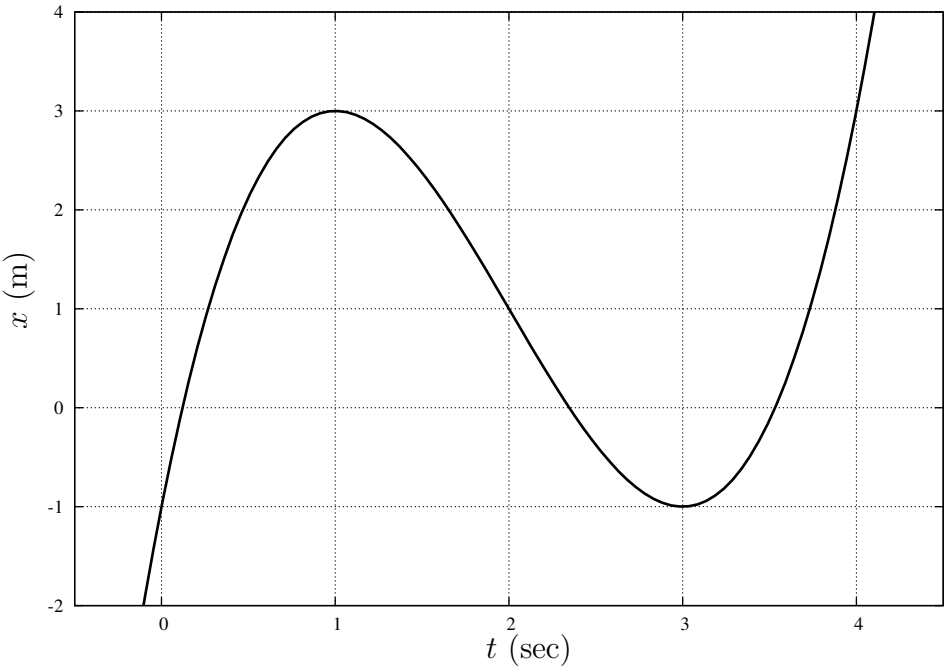


Figure 3.15: Problem 8

(c) Calculate analytically from $x = t^3 - 6t^2 + 9t - 1$ the location, velocity, and acceleration of the object at $t = 0$, 1, 2, 3, and 4 sec. Your answers to this part should be consistent with the plot (that is, with your answer to #8b).

(d) What is the object's displacement from 0 to 1 sec?

(e) What is the object's displacement from 1 to 2 sec?

(f) What is the object's average velocity from 0 to 4 sec?

(g) What is the object's average velocity at $t = 2$ sec?

(h) Obtain an approximate result for the instantaneous velocity of object at $t = 2$ sec directly from the graph. This should be consistent with your analytical result for the velocity at $t = 2$ sec.

(i) How many times in the interval $(0 \leq t \leq 4)$ is the object at rest?

(j) Is this problem lame, or what?

9. The bullet from a .357 magnum has a mass of 158 grains (10.2 g) and leaves the barrel, which is 6 inches (15 cm) long, at 440 m/s.
   (a) What is the acceleration of the bullet in the barrel? (Assume that this acceleration is constant, even though in reality this is far from true.)
   (b) How many g's is this?
   (c) How long does it take the bullet to travel through the barrel?
   (d) The bullet just happens to strike a life-size Barney doll and penetrates 20 cm as it comes to rest. Is the average number of g's of acceleration experienced by the bullet in the doll greater than, less than, or the same as that it underwent in the barrel? (You should be able to conclusively answer this without doing any numerical calculation, but your argument must still be rigorous and mathematical.)

10. A good school-bus driver can go from rest to 100 mph in 20 sec.
   (a) What is the bus's average acceleration?
   (b) How long does it take to reach 60 mph (assuming a constant acceleration)?

11. Suppose you have a tape measure and a stopwatch (or any equivalent means of measuring distances and times). How could you determine through observation and measurement whether the acceleration of a car was constant? And if the acceleration was constant, how could you determine the value of that acceleration? Be clear and precise about what you would measure, what relations you would use, and how you would use those measurements and relations: your explanation should not leave anything having to be inferred by reading between the lines.

12. (a) If you accelerate at 5.0 m/s^2 toward a cat standing in the road 100 m in front of you and it takes you 8.0 sec to reach the cat, with what velocity were you initially moving?
   (b) In what direction were you initially moving (that is, forward or backward)?
   (c) Describe your motion between the start of the problem and the time you reach the cat.

13. (a) If you accelerate at constant acceleration A toward a cat standing in the road a distance ℓ in front of you and it takes you time T to reach the cat, with what velocity were you initially moving?
   (b) i. For what critical value of T will $v_0 = 0$?
   ii. Analyze your result for your initial velocity when T is large or small.
   iii. What values of T, in terms of A and ℓ, can be considered large or small?

14. As you cruise along at 10 m/s, you see a 20 kg cat standing in the road 100 m ahead. In order to hit the cat within 5.0 sec, how quickly must you accelerate?

15. (a) As you cruise along at velocity V, you see a cat of mass m standing in the road a distance ℓ ahead. In order to hit the cat within a time T, how quickly must you accelerate?

    (b) At what velocity do you run over the cat?

    (c) Show that your result for the acceleration and for the velocity at which you run over the cat make sense when

        i. ℓ is very large.

        ii. T is very small.

        iii. V is very large or very small.

    (d) Show that your result for the acceleration and for the velocity at which you run over the cat make sense when

        i. ℓ is very small.

        ii. T is very large.

    If you find yourself perplexed and need a hint, see the footnote.[28]

16. As you cruise along at 15 m/s, you see a cat standing in the road 100 m ahead and hit the gas.

    (a) If you accelerate at 5.0 m/s^2, how long does it take you to reach the cat?

    (b) You should have found that mathematically there were two solutions for the time to reach the cat. How do you know which solution you want, and to what, physically, does the other solution correspond?

17. As you cruise along at velocity V, you see a cat standing in the road a distance ℓ ahead and hit the gas.

    (a) If you accelerate at acceleration A, how long does it take you to reach the cat?

    (b) You should have found that mathematically there were two solutions for the time to reach the cat. How do you know which solution you want, and to what, physically, does the other solution correspond?

[28] Think about the plot of velocity versus time, remembering that the acceleration in this situation is constant. What is the steepest plot of v versus t for which the acceleration is negative but the velocity remains nonnegative from $t = 0$ to $t = T$? To what value of the displacement ℓ does this plot correspond? And what is this telling you physically?

18. The velocity of some stupid object is given by

$$v = 3 + t^2$$

where v is in meters per second when t is in seconds.

When asked to determine the displacement of the object from $t = 0$ to $t = 4\,\text{sec}$, someone reasons as follows: at $t = 0$, $v = 3\,\text{m/s}$, and at $t = 4\,\text{sec}$, $v = 19\,\text{m/s}$. Therefore

$$x - x_0 = \frac{v + v_0}{2}\,t = \frac{19 + 3}{2}\,4 = 44\,\text{m}$$

(a) Why is this reasoning boneheaded?

(b) What is the correct result for the displacement?

19. A cat with a head start ℓ flees at a constant velocity v_c from a nuclear-powered steamroller. (That is, the cat is initially a distance ℓ in front of the steamroller.) The steamroller accelerates from rest at a constant acceleration a_r.

(a) Sketch the locations of the cat and steamroller as functions of time.

(b) How long does it take the steamroller to catch up with the cat?

(c) You should have found that mathematically there are two solutions for the time when the steamroller catches up with the cat. To what would the other solution correspond physically?

(d) Show that your solution for the time when the steamroller catches up with the cat makes sense in the following cases:

i. Large v_c.

ii. Large a_r.

iii. Small a_r.

(e) In the special case $\ell = 0$, how fast is the roller moving when it catches up with the cat? Make sense of your answer.

20. A school bus full of small children accelerates with constant acceleration from rest to speed $v_f = 100\,\text{mph}$ and then continues at a constant 100 mph until it reaches the school. The time for which the bus accelerates is equal to the time for which the bus travels at a constant 100 mph, and the total distance (distance traveled while accelerating plus distance traveled at constant speed) is ℓ.

    Determine the acceleration of the school bus and for how long it is accelerating. (The numerical value 100 mph is purely ornamental; you should not use it. Work in terms of variables.)

    It may help to think graphically. Or it may not. Who knows.

21. At an automotive testing facility, small children are tossed from blind alleyways in front of drivers. It is hypothesized that the drivers have a reaction time t_r, during which they continue at constant velocity, and that after that reaction time has elapsed and the brakes are applied, the vehicles experience a constant acceleration a_b.

    (a) If the stopping distance is ℓ when the initial velocity is V and 8ℓ when the initial velocity is $3V$, what are t_r and a_b?

    (b) If the stopping distance is 15ℓ when the initial velocity is $4V$, is this consistent with your results for the preceding part? If not, what's up with that?

22. A rebellious adolescent mass m moves along an x axis with velocity

$$v = 6t(1 - t)$$

    where v is in meters per second when t is in seconds.

    (a) Determine the acceleration of the mass as a function of time.

    (b) Determine the location of the mass as a function of time, given that the mass is at $x = 3\,\text{m}$ at $t = 0$.

    (c) Describe the motion from $t = 0$ to $t = 2\,\text{sec}$.

    (d) Determine the average velocity of the mass from $t = 0$ to $t = 2\,\text{sec}$.

    (e) Determine the average speed of the mass from $t = 0$ to $t = 2\,\text{sec}$.

23. It turns out that the motion of a mass m on the end of a vertical spring is given by

$$x(t) = A\cos\left(\sqrt{\frac{k}{m}}\,t\right)$$

where k is a positive constant parameter associated with the spring (the spring constant), A is a positive constant, and x represents the vertical displacement of the mass from the equilibrium point (that is, it represents how far the mass has moved above or below the point where it would naturally hang at rest).

(a) Determine the mass's velocity and acceleration as functions of time.

(b) Sketch $x(t)$, $v(t)$, and $a(t)$ and describe the motion.

(c) What significance does A have?

(d) What is the relation between the location x and acceleration a, and what does this mean physically?

24. A point mass m with a bad attitude moves along an x axis with acceleration

$$a = 24(1 - t^2)$$

where a is in meters per second per second when t is in seconds.

(a) Determine the velocity and location of the point mass as functions of time, given that at $t = 0$ the point mass is at $x = 6\,\text{m}$ and is moving at $16\,\text{m/s}$ in the negative x direction.

(b) Describe the motion.

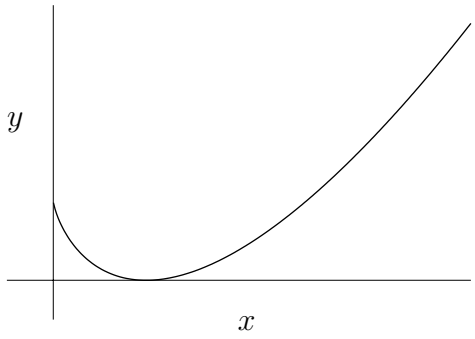


Figure 3.16: Problem 25

25. A point mass m that means well but has all sorts of emotional baggage moves along an x axis with velocity

$$v = \tilde{v} \ln \frac{t}{\tau}$$

where $\tilde{v}$ and τ are positive constants.

(a) What are the physical dimensions of $\tilde{v}$ and τ?

(b) Determine the acceleration and location of the point mass as functions of time, given that at $t = \tau$ the point mass is at the origin. If you've forgotten how to integrate logs, see the footnote.[29]

(c) Describe the motion and make sense of the relationships among $x(t)$, $v(t)$, and $a(t)$. If you need some help with $x(t)$, the plot of

$$y = x \ln x - x + 1$$

is shown in fig. (3.16).

26. Along an x axis the location of an object that lacks any real motivation in life is given as a function of the time t by

$$x = x_0 e^{-\beta t}$$

where x_0 and β are positive constants. Explain why each of the following observations is or is not sensible:

(a) The location of the object is $x = x_0$ at time $t = 0$.

(b) The location of the object is $x = x_0 e^{-\beta}$ at time $t = 1$.

(c) The location of the object $x \to 0$ as $t \to \infty$.

(d) The location of the object is $x = x_0/e$ at time $t = 1/\beta$.

[29] $\int dx \ \ln ax = x \ln ax - x$.

27. (A classic problem.) Tinker Bell gets caught inside a trash compactor. Initially, the sides of the trash compactor are a distance ℓ apart, and each side closes in on Tinker Bell at a constant speed v. In a panic, Tinker Bell, starting from one side of the compactor, flies continually from one side to the other at speed $v_{bell} > v$, instantaneously reversing direction each time she reaches the far side.

    (a) What total distance (without regard for direction) does Tinker Bell travel before getting squashed? If you can't see the quick way to work this out, see the footnote for a hint.[30]

    (b) This problem can also be solved by generating and then re-summing an infinite series. (A very similar technique of solving problems by series expansion, known as a perturbation series or perturbation expansion, is very common in physics.)

        i. When the sides of the compactor are their initial distance ℓ apart, how long does it take Tinker Bell to reach the opposite side?

        ii. How far has Tinker Bell traveled to reach the opposite side?

        iii. Once Tinker Bell has reached the opposite side, what is the new distance between the sides of the compactor?

        iv. Repeating these calculations for the next trip and the trip after that, you should be able to see a pattern and to express the total distance traveled by Tinker Bell before getting squashed as an infinite series. Determine this series.

        v. From a previous math course, you should be familiar with the result [31]

        $$\sum_{n=0}^{\infty} \epsilon^n = \frac{1}{1-\epsilon} \qquad (|\epsilon| < 1)$$

        Using this, you should be able to re-sum the series and simplify to get the same result as you did in #27a.

    (c) Consider *Zeno's paradox*: You shoot an arrow at a fleeing enemy. By the time the arrow gets to where the target was originally, the target has moved forward somewhat. And by the time the arrow gets to this new location of the target, the target has again moved forward somewhat. Since this process will repeat indefinitely, the arrow, it would seem, will never reach the target.

        Resolve this paradox.

        See the footnote if you need a hint.[32]

[30] Think in terms of distance = speed × time.

[31] If you have forgotten this result — or your previous math teacher fell victim to that pernicious insanity "less is more" — , you should be able to derive it on the spot by doing a Taylor expansion or simply by doing out the long division.

[32] How big are the successive time slices?

28. You accidentally let a water balloon fall from the window of your dorm room, which is 10 m above the ground.

    (a) If your roommate, who is 1.8 m tall, just happens to be standing directly below the window at that moment, how long will the balloon take to reach your roommate's head?

    (b) How fast will the balloon be moving when it reaches your roommate?

29. You accidentally let a water balloon fall from the window of your dorm room.

    (a) If your roommate just happens to be standing directly below the window at that moment, a vertical distance h below it, how long will the balloon take to reach your roommate?

    (b) How fast will the balloon be moving when it reaches your roommate?

30. If in $^{\#}28$ you hurl the balloon straight down at 10 m/s,

    (a) How long will it take to reach your roommate?

    (b) How fast will it be moving when it reaches your roommate?

    (c) What if you instead hurl the balloon up at 10 m/s: how long will it now take to reach your roommate, and at what speed will it hit him or her?

    (d) How, physically, does it make sense that you get the same speed on impact whether you throw the balloon up or down at 10 m/s?

31. If in $^{\#}29$ you hurl the balloon straight down at an initial *speed* $\tilde{v}$ ($\tilde{v} > 0$),

    (a) How long will it take to reach your roommate?

    (b) How fast will it be moving when it reaches your roommate?

    (c) What if you instead hurl it up at speed $\tilde{v}$: how long will it now take to reach your roommate, and at what speed will it hit him or her?

    (d) How, physically, does it make sense that you get the same speed on impact whether you throw the balloon up or down?

32. The water balloon in $^{\#}28$ is now released in such a way that it takes 1.0 sec to achieve its objective.

    (a) With what speed, and in what direction (up or down), was it thrown?

    (b) How, based on your results for the previous variations on this problem ($^{\#}28$ and $^{\#}30$), could you have reasoned out whether the balloon was thrown up or down without doing any numerical calculation at all?

33. The water balloon in #29 is now released in such a way that it takes a time T to achieve its objective.

    (a) With what initial velocity was it thrown?

    (b) For what range of values of T will this initial velocity be upward? Downward? Make sense of these ranges physically.

34. This problem requires thought, not calculation.

    (a) Can an object be at the origin and still have a velocity?

    (b) Can an object have zero velocity and still be accelerating? If so, try to think of an example.

    (c) How is #34a analogous mathematically to #34b?

    (d) You drop a container of pork fried rice on the floor. Is its acceleration greater during its descent through the air or as it hits the ground (that is, during that very brief time interval that it is smashing into the floor, in the process of coming to rest)?

35. In yet another bizarre accident, you lose your grip on a cat at the very moment that you are holding it over the mouth of a 200 m-deep abandoned mineshaft. (The numerical value 200 m for the depth h of the shaft is purely ornamental; you should instead work in terms of variables.)

    (a) At what speed does the cat strike the rocks below?

    (b) Does the cat land on its feet?

    (c) For how long is the cat falling?

    (d) What is the cat's speed when it has fallen halfway?

    (e) How long does the cat take to fall halfway?

    (f) Your answers to #35d and #35e are not half your answers to #35a and #35c. How does this make sense physically?

36. A piano of mass m falls from rest under the influence solely of the gravitational force mg. By reasoning or calculation, prove or disprove each of the following assertions.

    (a) Over equal distances, the piano's speed increases by equal amounts.

    (b) Over equal time intervals, the piano's speed increases by equal amounts.

    (c) The piano's speed at any given instant equals the total distance it has fallen divided by the total time it has been falling.

37. To see whether you might be interested in a career in psychology, you totally immobilize a younger sibling in a lawn chair by judicious application of duct tape, suspend a five-gallon bucket full of water from a tree limb directly over the sibling's head, and punch a small hole in the bottom of the bucket. Then you sit back to make observations as water drops ping the sibling's head, without respite, at maddeningly regular intervals. Bwahahahaha!

    It turns out that the timing of the drops is such that when a drop is just about to leave the bucket, four are in midair. If the distance between the two highest drops that are in midair is 1.0 m, how far is each of the airborne drops from the bucket?

38. Extensive research has shown that in order to stick, a half-chewed Gummy Bear must strike a ceiling at a minimal speed of 5.0 m/s.

    (a) If the ceiling of the common area in your dorm is 2.0 m above your release point when you make an underhand toss, at what minimal initial speed must you heave the Gummy Bear to get it to stick to the ceiling?

    (b) Suppose now that there were no ceiling to block the Gummy Bear's ascent.

        i. How high above the level of the ceiling would the Gummy Bear rise before falling back down?

        ii. How fast would the Gummy Bear be moving when it returned to the level from which it was thrown?

        iii. How long, from release to return, would the Gummy Bear be in the air?

39. A porcupine is dropped from rest from the edge of a cliff. A time T later a water balloon is hurled vertically downward at speed $\tilde{v}$.

    (a) How far below the top of the cliff, in terms of time and distance, will the balloon collide with the porcupine?

    (b) Make physical sense of your solutions for the time and distance in the limit of large $\tilde{v}$.

    (c) Why is the critical value $\tilde{v} = gT$? How does this make sense physically?

40. A cat, coated with the material used to make "bouncy balls," is dropped from rest down a mine shaft of depth $h > 0$. Upon impact with the bottom of the shaft, the cat's velocity reverses direction. You want to drop a sack full of blasting caps from rest with just the right timing to reach the cat, as it rises back up, at the midpoint of the shaft. How much time should be allowed to elapse between dropping the cat and dropping the blasting caps?

41. Knowing that the height of the smaller of two buildings that stand side by side is h, you want to determine the height of the taller building, and you want to do this in the most difficult, perverse way you can think of. So you get a stopwatch and fire a cat vertically upward out of a homemade cannon. You note that it takes the cat a time T to go from the roof line of the smaller building to the roof line of the taller building. You also note that it takes the cat a time T' to rise from and return to the roof line of the taller building (having reached its apex and turned around in between). It then takes the cat a time T to fall from the roof line of the taller building to the roof line of the smaller building. From this information, determine the height of the taller building.

42. Elvis is spotted at a convenience store located at coordinates $(x, y) = (10\,\text{mi}, 20\,\text{mi})$. Space aliens then kidnap Elvis and drive him 90 mi away, traveling at an angle of 210° with the positive x axis, in a pink Cadillac with heart-shaped windows, a fake leopard-skin steering-wheel cover, fuzzy dice hanging from the mirror, and one of those little cardboard pine-scented Christmas trees. The journey takes $1\frac{1}{2}$ hr.

    (a) What are the components of Elvis's average velocity?

    (b) What is Elvis's average speed?

    (c) What are Elvis's final coordinates?

43. On a beautiful sunny day at the shore, you no sooner get your blanket spread out, set up your beach umbrella, put on your suntan lotion, and lean back in your beach chair with a cool beverage than suddenly dark clouds appear out of nowhere and it starts pouring. Being unusually perceptive, you notice that the rain, which is coming straight down when you are at rest, pelts your face at angle θ to the vertical as you are dashing at speed V, over an essentially level stretch of beach, back to the car.

    (a) At what vertical speed is the rain falling?

    (b) At what speed is the rain pelting your face as you dash back to the car?

44. An object from a dysfunctional family moves such that its position vector $\mathbf{r}$ as a function of the time t is

$$\mathbf{r} = At\,\hat{\mathbf{x}} + B\,\hat{\mathbf{y}} + C\sin\omega t\,\hat{\mathbf{z}}$$

    where A, B, C, and ω are all positive constants.

    (a) What are the physical dimensions of A, B, C, and ω?

    (b) Determine the velocity and acceleration of the object as functions of time.

    (c) What does the trajectory look like?

    (d) Is it possible that this object could recover with therapy?

45. The acceleration of an object suffering from low self esteem is given by

$$\mathbf{a} = Ae^{-\beta t}\,\hat{\mathbf{z}}$$

where A and β are positive constants and t is the time.

(a) If at $t = 0$ the object is at the point (x_0, y_0, z_0) and is moving at speed v_0 in the xy plane at 45° to the positive x and positive y axes, determine the object's velocity and location as functions of time.

(b) What does the trajectory look like?

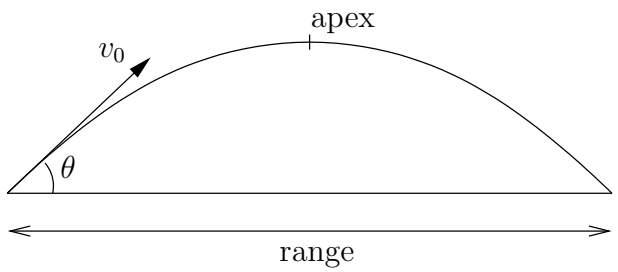


Figure 3.17: Problem 46

46. Fig. (3.17) shows an object launched at angle θ at initial speed v_0 over level ground.

(a) Show that the height of the apex (that is, the highest point) above the ground is

$$h = \frac{v_0^2 \sin^2\theta}{2g}$$

(b) Show that the range is

$$R = \frac{v_0^2 \sin 2\theta}{g}$$

See the footnote if you need a hint.[33]

(This is of course neglecting air resistance. For a calculation of range with air resistance, see Appendix A.)

(c) For a fixed initial velocity v_0, what launch angle θ will maximize the range?

(d) The trajectory is drawn symmetrically in fig. (3.17): the ascent is the mirror image of the descent. Explain how you can see from the relations governing projectile motion that this must be so.

[33] What is the overall vertical displacement? Also, remember that $\sin 2\theta = 2\sin\theta\cos\theta$.

47. A water balloon is fired out of a water-winger at 30° to the vertical at an exhilarating 100 m/s. The water-winger is at the edge of a vertical cliff 100 m above a level plain.

    (a) How long is the balloon in the air? That is, what is the time of flight?

    (b) You should have found that mathematically there were two solutions for the time in the preceding part. To what would the other solution correspond physically?

    (c) How far from the base of the cliff does the balloon land?

    (d) How long does it take the balloon to reach its highest point above the plain?

    (e) How high above the plain is this highest point?

    (f) What are the magnitude and direction of the balloon's velocity at this highest point?

    (g) What are the magnitude and direction of the balloon's velocity when it lands?

48. A water balloon is fired out of a water-winger at initial velocity V at angle φ to the vertical. The water-winger is at the edge of a vertical cliff of height h above a level plain.

    (a) How long is the balloon in the air? That is, what is the time of flight?

    (b) You should have found that mathematically there were two solutions for the time in the preceding part. To what would the other solution correspond physically?

    (c) How far from the base of the cliff does the balloon land?

    (d) How long does it take the balloon to reach its highest point above the plain?

    (e) How high above the plain is this highest point?

    (f) What are the magnitude and direction of the balloon's velocity at this highest point?

    (g) What are the magnitude and direction of the balloon's velocity when it lands?

49. If instead the balloon in #47 is fired with an initial velocity such that it lands 60 m from the base of the cliff 6.0 sec after being fired, what are the magnitude and direction of the balloon's initial velocity?

50. (a) If instead the balloon in #48 is fired with an initial velocity such that it lands a horizontal distance ℓ from the base of the cliff a time T after being fired, what are the magnitude and direction of the balloon's initial velocity?

    (b) Make physical sense of your results when

        i. T is large.

        ii. T is small.

51. You buy a 30-06 to protect your stash of ramen noodles in your dorm room. As a crude way of determining its muzzle velocity, you get a friend to stand with an apple on his or her head a level 500 yd (457 m) away. When the gun is fired horizontally, it turns out that the bullet drops 5.12 ft (1.56 m) on the way to the target and strikes your friend in the shin.

    What is the muzzle velocity of the rifle? (That is, what is the speed of the bullet as it leaves the barrel?) As usual, neglect air resistance.[34]

52. You are zipping along a zip line, on your way to doing some cool, Mission Impossible-like thing of the sort that you'd expect of someone who uses a zip line, when your keys fall out of your pocket in an embarrassingly dorky way. The zip line is very very nearly horizontal, so that your 30 m/s speed remains very nearly constant, and you are 20 m above the ground. As usual, we will ignore air resistance.

    (a) How long does it take the keys to reach the ground?

    (b) In what direction do you have to turn your head as you watch your keys fall?

    (c) Is it reasonable to have ignored air resistance in this problem? If not, how would your answers to the previous parts be affected by air resistance?

[34] Air resistance is actually very significant for such high-velocity projectiles in spite of their small size and rounded shape. In fact, a 180 grain soft-point has a muzzle velocity of 810 m/s and would drop 73.7 inches at 500 yd even when zeroed in at 100 yd — fired horizontally, the drop would be between 7 and 8 ft.

53. You are zipping along a zip line, on your way to doing some cool, Mission Impossible-like thing of the sort that you'd expect of someone who uses a zip line, when your keys fall out of your pocket in an embarrassingly dorky way. The zip line is very very nearly horizontal, so that your speed V remains very nearly constant, and you are a height h above the ground. As usual, we will ignore air resistance.

    (a) How long does it take the keys to reach the ground?

    (b) In what direction do you have to turn your head as you watch your keys fall?

    (c) Is it reasonable to have ignored air resistance in this problem? If not, how would your answers to the previous parts be affected by air resistance?

54. While cleaning out the gutters, you fall on your butt and skid down a section of roof tilted at 15° to the horizontal. You skid off the edge of the roof, which is 6.0 m vertically above the ground, at 2.0 m/s.

    (a) How far do you travel horizontally (after leaving the roof) on your way down to the ground?

    (b) At what speed do you hit the ground?

55. Inspired by six bowls of Chocolate Frosted Sugar Bombs and several hours of cartoons one Saturday morning, you attempt to jump the Grand Canyon on your bicycle. The two sides of the canyon are 1600 m apart but at essentially the same height. If you pedal off of a small ramp inclined at 30° to the horizontal, at what speed would you have to have made the jump in order to have made it to the other side?

56. Inspired by six bowls of Chocolate Frosted Sugar Bombs and several hours of cartoons one Saturday morning, you attempt to jump the Grand Canyon on your bicycle. The two sides of the canyon are a distance ℓ apart but at essentially the same height. If you pedal off of a small ramp inclined at angle θ to the horizontal, at what speed would you have to have made the jump in order to have made it to the other side?

57. Walking across campus one beautiful spring day you see your boy/girl friend, 30 m away, back to you, flirting with someone else. You decide to handle the situation with maturity and dignity and hurl a 10-lb calculus book at him/her.[35] When you heave the book at 15° to the conveniently level ground, it falls 5.0 m short. At what angle should you heave the rest of your knapsack to get satisfaction?

[35] Who said math wasn't good for anything?

58. (Based on *Monty Python & the Holy Grail*.) You attempt to catapult a Trojan rabbit over a castle wall. The rabbit leaves the catapult at 44 m/s, and the target lies a horizontal distance of 100 m away over level ground.

    (a) At what angle to the horizontal should the catapult be fired in order to hit the target?

    (b) You should have found that for the angle there are two solutions that are complements of each other. Describe the differences between the trajectories yielded by these two angles.

    (c) The 10 m high castle wall stands a horizontal distance of 30 m from the catapult. For the shallower angle of fire, will the rabbit clear the wall?

    (d) Is the rabbit on the way up or down when it hits the wall? How do you know this?

59. (Based on *Monty Python & the Holy Grail*.) You attempt to catapult a Trojan rabbit over a castle wall. The rabbit leaves the catapult at speed V, and the target lies a horizontal distance R away over level ground.

    (a) At what angle θ to the horizontal should the catapult be fired in order to hit the target?

    (b) You should have found that there are two solutions for θ. Describe the differences between the trajectories yielded by these two angles.

    (c) The castle wall, height h, stands a horizontal distance ℓ from the catapult. If the two solutions for the angle θ just happen (in order to keep the calculation simple) to be equal, how high can the castle wall be and still be cleared by the rabbit?

    (d) Suppose that the castle wall in the preceding part is too high. How could you ascertain whether the rabbit was on the way up or down when it struck the wall? Without actually doing it, describe the calculation you would do and the algebraic condition you would test.

60. You are assigned the task[36] of generating exercises in projectile motion. If you approach this task in the most cynical possible way and reckon from the numbers of physical variables and relations among them, how many distinct types of projectile-motion problems are there?

[36] Don't say *are tasked with*. *Task* is not a verb. What is wrong with your generation?

61. A cinder block and a large pumpkin are launched from skeet traps on opposite sides of the Grand Canyon, which are at the same height and a distance ℓ apart. The block is launched horizontally at speed $v_0/\sqrt{2}$, the pumpkin at initial speed v_0 at an angle of $\frac{\pi}{4}$ above the horizontal. By amazing good luck, it happens that the block and pumpkin collide in midair.

    (a) Make a rough sketch of the trajectories of the block and the pumpkin.

    (b) For a collision to occur, the block and the pumpkin must have been launched at different times. Which was launched first? You should be able to reason this out without doing any calculation.

    (c) Okay, now you have to calculate: determine the time difference between the launching of the block and the launching of the pumpkin. The algebra is not bad if you approach it circumspectly.

62. You threaten a younger sibling standing beside you with a water balloon, causing the sibling to flee from you at a constant speed v_s across the usual conveniently level ground.

    (a) If you throw the balloon at an initial speed v_0 at the same moment that the sibling starts fleeing, at what angle do you need to throw the balloon for a money shot?

    (b) What if the sibling, instead of running over level ground, is running up an incline at angle φ to the horizontal?

63. You are fighting an uphill battle, assaulting entrenched enemy positions of the neighboring dorm a distance ℓ farther up a straight hillside inclined at angle θ to the horizontal. You (mass m_y) can throw a water balloon (mass m_b) at speed V.

    (a) Set up, *but do not actually solve*, a relation or relations that could be solved for the angle φ *to the horizontal* at which should you toss the water balloon in order to hit the enemy. See the footnote if you need a hint.[37]

    (b) At what angle to the incline should you toss the water balloon in order to maximize the range? See the footnote if you are really bad at trig relations.[38]

[37] Recall that when tossing over level ground, the condition that you returned to the ground was $y - y_0 = 0$. Think about what condition corresponds to returning to the ground in the sense of landing on the hillside. Alternatively, it is possible to work this problem out using a tilted coordinate system — just be careful, if you do it this latter way, about how you include the acceleration due to gravity.

[38] Recall that $\tan\alpha = \cot\beta$ means that $\alpha = \frac{\pi}{2} - \beta$.

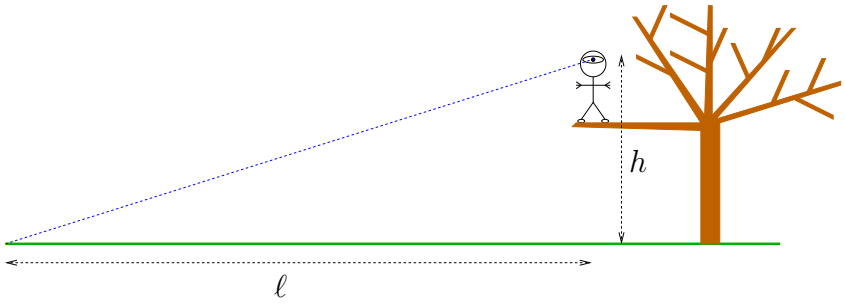


Figure 3.18: Problem 64

64. Fig. (3.18) shows a variation on the classic and very politically incorrect monkey-hunter problem: you go hunting for tree-dwelling cyclopes.[39] You spot one a horizontal distance ℓ away up on a tree branch at height h above the ground. You sight the barrel of your black-powder, flintlock cyclops gun directly along the line of sight to the target (the blue dashed line in fig. (3.18)) — forgetting to take into account the drop due to gravity. Realizing this, the cyclops, upon seeing the muzzle flash, drops from rest from the branch at the instant you fire.

    (a) Prove that the cyclops is dead meat.

    (b) What assumptions, if any, did you need to make about the muzzle velocity v_0 of your cyclops gun?

65. Determine whether each of the following three-dimensional cases is physically possible.

    If the case is physically possible, describe the motion that will occur and try to think of a realistic example of such motion.

    (a) The velocity and acceleration are initially both zero.

    (b) The velocity is initially zero, but the acceleration is nonzero.

    (c) The acceleration is zero, but the velocity is initially nonzero.

    (d) Both the velocity and the acceleration are nonzero, and

        i. The acceleration is in the same direction as the velocity.

        ii. The acceleration is opposite in direction (that is, antiparallel) to the velocity.

        iii. The acceleration is always perpendicular to the velocity.

        iv. The acceleration has components both parallel (or antiparallel) and perpendicular to the velocity.

[39] Dude: *cyclopes* is the plural of *cyclops*.

66. You fritter away your youth by repeatedly tossing a golf ball straight up into the air and catching it.

    (a) Would the behavior of the ball differ if, instead of leaning idly against a lamppost and chewing a toothpick, you were doing this on a train moving at constant velocity? If so, precisely how would it differ?

    (b) What would the trajectory of the ball look like to someone standing at rest on the platform as the train passed a station?

    (c) Determine whether and precisely how the behavior of the ball would differ if, instead of moving at constant velocity, the train were

        i. Speeding up.

        ii. Slowing down.

        iii. Making a turn to the right.

        For definiteness, assume that in these cases you are facing toward the front of the train as you toss the golf ball.

    (d) Determine whether and precisely how the behavior of the ball would differ if, instead of leaning idly against a lamppost and chewing a toothpick, you were doing this in an elevator

        i. Starting up toward a higher floor.

        ii. Starting down toward a lower floor.

        iii. Coming to a stop as it arrives at a higher floor.

        iv. Coming to a stop as it arrives at a lower floor.

67. (a) Suppose you take a picture, at some random instant while it is airborne, of a grenade that your friend tosses straight up. Where in the flight of the grenade is your picture most likely to catch the grenade? Note that *random instant while it is airborne* means at any instant of time, with equal probability, while the grenade is in the air.

    (b) Suppose that your friend instead tosses the grenade in the more usual fashion, with some horizontal as well as upward initial velocity. Where in its flight is your picture most likely to catch the grenade?

    (c) Even though you are still in one of the outer layers of The Great Onion of Understanding and we have not yet studied the motion of pendula, but you've certainly all seen one swinging back and forth, and once you've seen one, you've seen them all. If you take a picture of a pendulum at some random instant, where in its swing is your picture most likely to catch it?

68. Suppose you run in circles of 12 m radius at 10 mph = 4.5 m/s while screaming your head off.[40]

    (a) Determine the frequency, angular frequency, period, tangential velocity, and centripetal acceleration of your circular motion.

    (b) Is this centripetal acceleration perceptible?

69. The radius of the Earth's nearly circular orbit around the Sun is 1.5×10^{11} m. If we take a year to be an even 365 days,

    (a) At what speed is the Earth orbiting around the Sun?

    (b) Is this speed of orbit perceptible?

    (c) What is the Earth's centripetal acceleration toward the Sun?

    (d) Is this centripetal acceleration perceptible?

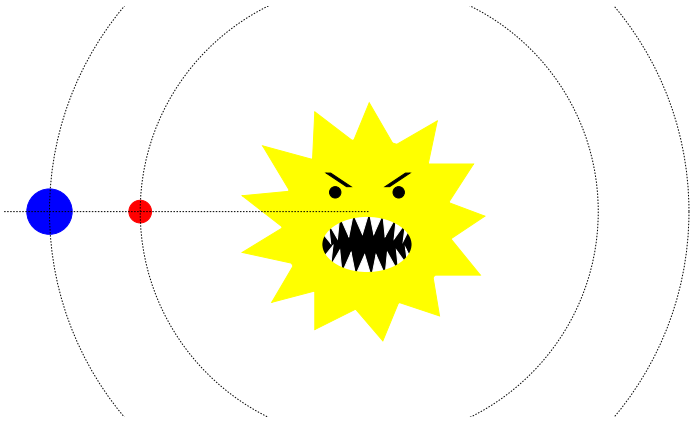


Figure 3.19: Problem 70

70. Astronomically, two (or more) planets are said to be in syzygy when they lie along the same line through the Sun, as shown in fig. (3.19) for the Earth (blue) and Venus (red). If the orbital periods of the Earth and Venus are 365.24219 and 224.70069 days, respectively, how many Earth years elapse between successive syzygies of Earth and Venus? See the footnote if you need a hint.[41]

[40] Remember,

When in panic,
When in doubt,
Run in circles,
Scream and shout!

[41] You want to think in terms of angular velocities when figuring out how long it will take one planet to "lap" the other.

71. In a daring experiment, you turn the key in the ignition of your suburban assault vehicle, wave to the dean of students, rev the engine, and start doing donuts in the parking lot. With the wheels turned so that you travel in a circle of radius 12 m, you put the pedal to the metal and increase your speed at a constant rate as you go from rest to 100 mph = 45 m/s in 15 sec.

    (a) Determine your tangential, angular, and centripetal accelerations when you have reached 45 m/s. Which of these accelerations is constant throughout your motion and which are not?

    (b) How many donuts (that is, revolutions) do you make during the 15 sec that you are accelerating?

    (c) How long (that is, how much time) does it take you to reach a centripetal acceleration of 1 g?

    (d) How many g's are you experiencing when you have reached this point? (Careful: The number of g's you experience is determined by your *net* (total) acceleration.)

    (e) Which of your numerical results for the above parts are realistic and which unrealistic?

72. An object ill at ease at social functions revolves in a circle of radius ρ, centered at the origin, in the xy plane. At time $t = 0$ the object starts from rest on the positive x axis, and it experiences an angular acceleration $\alpha = At$, where A is a positive constant.

    (a) Describe the motion.

    (b) How long does it take the object to complete its first full revolution?

    (c) What is the object's speed upon completing its first full revolution?

    (d) What are the object's tangential and centripetal accelerations upon completing its first full revolution?

73. (The classic cycloid problem.) You roll a wheel of radius ρ along level ground.

    (a) Convince yourself that, if you set up your xy axes so that $y = 0$ is ground level and the center of the wheel starts at $x = 0$ at time $t = 0$, and if the wheel is moving in the positive x direction at a constant speed v, then the trajectory of the point P that starts at the top of the wheel at $t = 0$ is given by

    $$x = \rho(\omega t + \sin \omega t) \qquad\qquad y = \rho(1 + \cos \omega t)$$

    where $\omega = v/\rho$ is the angular frequency of the wheel's rotation. See the footnote if you need a hint.[42]

    (b) Determine the components of the velocity and acceleration of point P as functions of time.

    (c) From these components, determine the magnitude and direction of the velocity and acceleration of the point P when it is at

    i. The top of the wheel.
    ii. The bottom of the wheel.
    iii. The midpoint of the wheel on the front side.
    iv. The midpoint of the wheel on the back side.

    See the footnote if you need a hint.[43]

    (d) Sketch the trajectory of the point P.

74. (a) Determine the trajectory traced out by the position vector

    $$\mathbf{r} = \mathbf{r}_0 + \mathbf{r}_1 t \tag{3.43}$$

    where $\mathbf{r}_0$ and $\mathbf{r}_1$ are constants and t is the time.

    (b) Determine the trajectory traced out by the position vector

    $$\mathbf{r} = \mathbf{r}_0 + \mathbf{r}_1 t^2 \tag{3.44}$$

    where $\mathbf{r}_0$ and $\mathbf{r}_1$ are constants and t is the time.

    (c) What is the difference in the motions described by eqq. (3.43) and (3.44)?

[42] Note that all points on the wheel share the forward velocity v, and that superposed on top of this forward motion is a pure rotation. Since there is no slipping or skidding, the net velocity of the point of contact with the ground, which is the sum of that point's forward velocity v and its tangential rotational velocity, must vanish. (*Vanish* of course being big-people speak for *equal zero*.)

[43] Remember that ωt represents the angle through which the wheel has turned, and this has simple values at the four points in question.

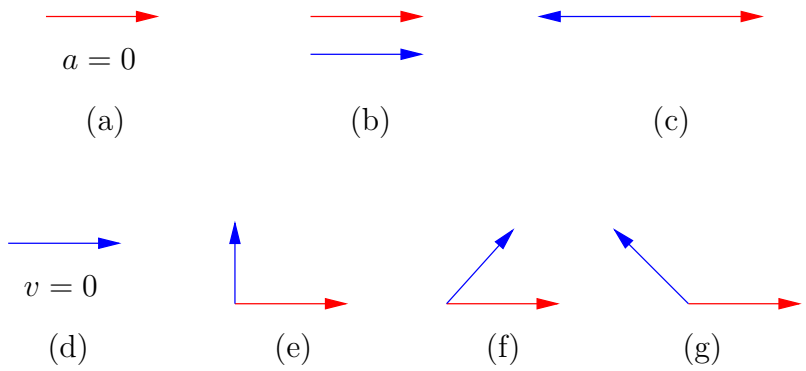


Figure 3.20: Problem 75

75. The red and blue arrows in fig. (3.20) represent the velocity and acceleration of an object, respectively. For each of the cases in fig. (3.20) describe how the speed and direction of motion of the object is changing at that instant.

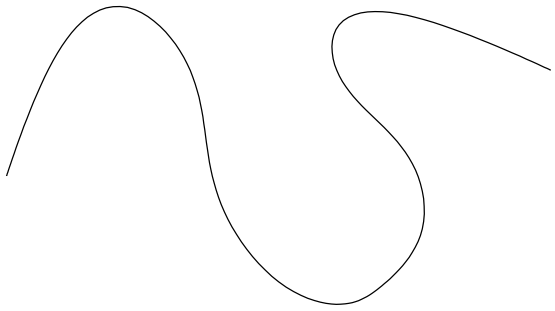


Figure 3.21: Problem 76

76. Fig. (3.21) shows the trajectory of a desperately lost object far out in the lonely void of space. Is the object experiencing an acceleration?

77. Determine the trajectory traced out by the position vector

$$\mathbf{r} = \rho \cos \omega t \, \hat{\mathbf{x}} + \rho \sin \omega t \, \hat{\mathbf{y}} + \beta t \, \hat{\mathbf{z}}$$

where ρ, ω, and β are constants and t is the time.

78. The position vector of an object having trouble figuring out what to do with its future is given as a function of the time t by

$$\mathbf{r} = A\theta\,\hat{\mathbf{r}} \qquad \text{with} \qquad \theta = \omega t$$

where A and ω are positive constants.

(a) What are the physical dimensions of A?

(b) Describe the object's trajectory.

(c) Determine, in polar coordinates, the object's

i. Velocity.

ii. Acceleration.

(d) Physically interpret the terms in your results for the velocity and acceleration.

79. The position vector of a vaguely anxious object is given as a function of the time t by

$$\mathbf{r} = A\,\hat{\mathbf{r}} \qquad \text{with} \qquad \theta = \beta t^2$$

where A and β are positive constants.

(a) What are the physical dimensions of A and β?

(b) Describe the object's trajectory.

(c) Determine, in polar coordinates, the object's

i. Velocity.

ii. Acceleration.

(d) Physically interpret the terms in your results for the velocity and acceleration.

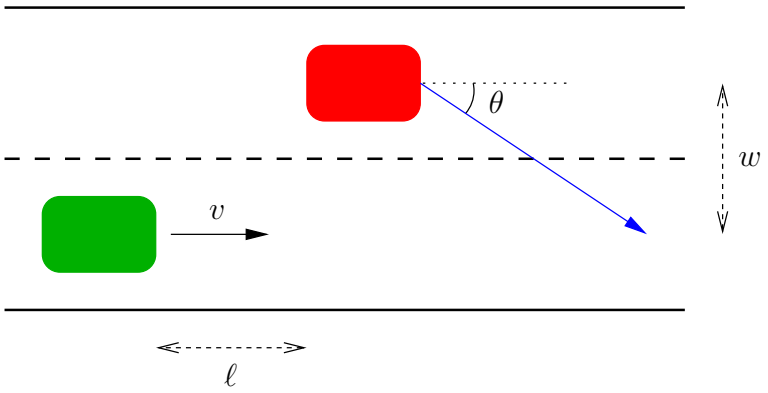


Figure 3.22: Problem 80

80. To save maybe 3 sec at the outside on your daily commute, at the risk of taking 50 yr off your life, you cut in front of another driver. To make the lane change, you maintain a constant speed but jerk the wheel so that you suddenly change your direction of motion by angle θ, as shown in fig. (3.22). The lanes are of width w, and at the instant you jerk the wheel you have a head start ℓ on the driver in the adjacent lane.

    (a) At what minimal speed must you be traveling to avoid a collision?

    (b) Make physical sense of the dependence of your answer on w, v, ℓ, and θ.

81. Your rowing speed in still water is 4 mph. You attempt to row straight across a river 2 mi wide, but a 3 mph current carries you downstream.

    (a) What is your speed relative to the shore?

    (b) In what direction are you actually moving? That is, at what angle relative to the shore?

    (c) How long does it take you to reach the other side of the river?

    (d) How far will you have gone downstream by the time you reach the other side?

    (e) At what angle to the shore would you have to row in order to travel directly (that is, perpendicularly) across the river?

    (f) How long would it take you to get across the river rowing at this angle?

82. Suppose a commercial passenger plane's airspeed (the speed at which the plane travels relative to the air) is v but that, to counter the windstorm produced by a nuclear blast, it has to alter its direction of flight by 60° and that its resulting velocity relative to the ground is $\frac{3}{2}v$. Determine the possible speeds and directions of the wind.

83. Suppose a plane's airspeed (the speed at which the plane travels relative to the air) is v and that the plane is to make a round trip, traveling distance ℓ each way.[44]

    (a) Determine the time T for the round trip if there is no wind.

    (b) Determine the time for the round trip if the wind, blowing at speed v_w, is a tailwind on the way out and a headwind on the way back. Express your answer in terms of T and the speeds, in a respectably simplified form.

    (c) Determine the time for the round trip if the wind, blowing at speed v_w, is a perpendicular crosswind both on the way out and the way back. Express your answer in terms of T and the speeds, in a respectably simplified form.

    (d) How do the times for these three cases compare? Which is longest, and which shortest?

84. You are cycling[45] at 20 mph down a road along which the traffic, moving at a uniform 40 mph, passes a bystander with equal frequency in either direction. From your perspective, at what rate are oncoming cars passing you relative to the rate at which you are being passed by cars from behind?

[44] The calculations in this problem are similar to those of the famous Michelson-Morley experiment, which attempted to detect the effect of the Earth's motion through a hypothesized æther on the speed of light by comparing the timing of light beams traveling two paths, one parallel to the Earth's orbital velocity and the other perpendicular to it. Here the plane and wind are analogous to the light and the æther (or, more precisely, the motion through the æther), respectively.

[45] *Cycling* means pretty much the same thing as *riding a bike*, except that it's done by people wearing ridiculously expensive gear who are just *so* much more awesome than mere mortals.

3.5 Sketchy Answers

(2a) 30 mph.

(2c) 4 min.

(2d) 48 mph.

(3a) 8:08:45.

(3b) 8.0 ft/sec.

(3c) 6.4 ft/sec.

(3d) 10 ft/sec.

(4(a)i) $\dfrac{2v_1v_2}{v_1+v_2}$.

(4(b)i) $\frac{1}{2}(v_1+v_2)$.

(7a) -80 m.

(7c) $\frac{25}{3}$ m/s.

(8c) The *combined* set of numerical values of for x (in m), v (in m/s), and a (in m/s^2) are, all jumbled up together, $\{-12, -6, -3, -1, 0, 1, 3, 6, 9, 12\}$.

(8d) 4 m.

(8e) -2 m.

(8f) 1 m/s.

(8j) We've seen worse. But not much.

(9a) 6.5×10^5 m/s^2.

(9b) 6.6×10^4.

(9c) 6.8×10^{-4} sec.

(10a) 5.0 mph/sec.

(10b) 12 sec.

(12a) The *speed* was 7.5 m/s.

(13a) $\dfrac{\ell - \frac{1}{2}AT^2}{T}$.

(14) 4.0 m/s^2.

(15a) $\dfrac{2(\ell - VT)}{T^2}$.

(15b) $\dfrac{2\ell}{T} - V$

(16) 4.0 sec or -10 sec.

(17) The two solutions are $\dfrac{-V \pm \sqrt{V^2 + 2A\ell}}{A}$.

(18b) $\frac{100}{3}$ m.

(19b) $\dfrac{v_c + \sqrt{v_c^2 + 2a_r\ell}}{a_r}$.

(19e) $2v_c$.

(20) $a = \dfrac{3v_f^2}{2\ell}, \quad t = \dfrac{2\ell}{3v_f}$.

(21a) $t_r = \dfrac{\ell}{6V}, \quad a_b = \dfrac{3V^2}{5\ell}$.

(22a) $6(1 - 2t)$.

(22b) $3 + t^2(3 - 2t)$.

(22d) -2 m/s.

(22e) 3 m/s.

(23a) $v = -A\sqrt{\dfrac{k}{m}}\sin\left(\sqrt{\dfrac{k}{m}}\,t\right), \quad a = -\dfrac{Ak}{m}\cos\left(\sqrt{\dfrac{k}{m}}\,t\right)$.

(23d) $a = -\dfrac{k}{m}\,x$.

(24a) $v = -16 + 24t - 8t^3$ and $x = 6 - 16t + 12t^2 - 2t^4$.

(25b) $a = \dfrac{\tilde{v}}{t}$ and $x = \tilde{v}\Big(t\ln\frac{t}{\tau} - (t - \tau)\Big)$.

(27(b)i) $\dfrac{\ell}{v + v_{bell}}$.

(27(b)ii) $\dfrac{v_{bell}}{v + v_{bell}}\,\ell$.

(27(b)iii) $\dfrac{v_{bell} - v}{v_{bell} + v}\,\ell$.

(27(b)iv) $\dfrac{v_{bell}}{v + v_{bell}}\,\ell\sum_{n=0}^{\infty}\left(\dfrac{v_{bell} - v}{v_{bell} + v}\right)^n$.

(28a) 1.3 sec.

(28b) 13 m/s.

(29a) $\sqrt{\dfrac{2h}{g}}$.

(29b) The *speed* will be $\sqrt{2gh}$.

(30a) 0.63 sec.

(30b) 16 m/s.

(30c) 2.7 sec, 16 m/s.

(31a) $\dfrac{-\tilde{v} + \sqrt{\tilde{v}^2 + 2gh}}{g}$.

(31b) The *speed* will be $\sqrt{\tilde{v}^2 + 2gh}$.

(31c) $t = \dfrac{\tilde{v} + \sqrt{\tilde{v}^2 + 2gh}}{g}$ and a speed of $\sqrt{\tilde{v}^2 + 2gh}$.

(32a) The *speed* was 3.3 m/s.

(33a) $\dfrac{\frac{1}{2}gT^2 - h}{T}$.

(35a) If you evaluate your result numerically, it should yield 63 m/s.

(35b) It doesn't matter.

(35c) 6.4 sec.

(35d) 44 m/s.

(35e) 4.5 sec.

(37) $\frac{1}{3}$, $\frac{4}{3}$, 3, and $\frac{16}{3}$ m.

(38a) 8.0 m/s.

(38(b)i) 1.3 m.

(38(b)iii) 1.6 sec.

(39a) Time $\dfrac{gT^2}{2(\tilde{v} - gT)}$, distance $\frac{1}{2}gT^2\left(\dfrac{\tilde{v} - \frac{1}{2}gT}{\tilde{v} - gT}\right)^2$.

(40) $2(\sqrt{2} - 1)\sqrt{\dfrac{h}{g}}$.

(41) $h + \frac{1}{2}gT(T + T')$.

(42a) $-30\sqrt{3}$ mph, -30 mph.

(42b) 60 mph.

(42c) $(10 - 45\sqrt{3}\text{ mi}, -25\text{ mi})$.

(43a) $V \cot\theta$.

(43b) $V \csc\theta$.

(44b) $\mathbf{v} = A\,\hat{\mathbf{x}} + \omega C \cos\omega t\,\hat{\mathbf{z}}, \quad \mathbf{a} = -\omega^2 C \sin\omega t\,\hat{\mathbf{z}}$.

(45a) $\mathbf{v} = \dfrac{v_0}{\sqrt{2}}(\widehat{\mathbf{x}} + \widehat{\mathbf{y}}) + \dfrac{1}{\beta}A\big(1 - e^{-\beta t}\big)\,\widehat{\mathbf{z}}$

$\mathbf{r} = \left(x_0 + \dfrac{v_0}{\sqrt{2}}\,t\right)\widehat{\mathbf{x}} + \left(y_0 + \dfrac{v_0}{\sqrt{2}}\,t\right)\widehat{\mathbf{y}} + \left(z_0 + \dfrac{1}{\beta}A\,t - \dfrac{1}{\beta^2}A\big(1 - e^{-\beta t}\big)\right)\widehat{\mathbf{z}}$

(47a) 19 sec.

(47c) 940 m.

(47d) 8.8 sec.

(47e) 480 m.

(47g) 110 m/s at 63°.

(48a) $\dfrac{V\cos\varphi + \sqrt{(V\cos\varphi)^2 + 2gh}}{g}$.

(48c) $\dfrac{V\sin\varphi}{g}\,(V\cos\varphi + \sqrt{(V\cos\varphi)^2 + 2gh})$.

(48d) $\dfrac{V\cos\varphi}{g}$.

(48e) $h + \dfrac{V^2\cos^2\varphi}{2g}$.

(48g) $v = \sqrt{V^2 + 2gh}$ with the angle to the horizontal given by

$\tan\theta = -\dfrac{\sqrt{(V\cos\varphi)^2 + 2gh}}{V\sin\varphi}$

(49) 16 m/s at 52°.

(50) $v_0 = \dfrac{1}{T}\sqrt{\ell^2 + (\frac{1}{2}gT^2 - h)^2}$ with the direction given by

$\cos\theta = \dfrac{\ell}{\sqrt{\ell^2 + (\frac{1}{2}gT^2 - h)^2}}$

(51) 810 m/s. Didn't you read the footnote? Sheesh.

(52a) 2.0 sec.

(53a) $\sqrt{\dfrac{2h}{g}}$.

(54a) 2.0 m.

(54b) 11 m/s.

(55) 135 m/s.

(56) $v_0 = \sqrt{\dfrac{g\ell}{\sin 2\theta}}$.

(57) 18°.

(58a) 15° or 75°.

(58c) It strikes the wall at a height of 5.7 m.

(59a) If $\varphi = \dfrac{1}{2}\arcsin\dfrac{gR}{V^2}$ then $\theta = \varphi$ or $\theta = \frac{\pi}{2} - \varphi$.

(59c) $h < \ell(1 - \dfrac{g\ell}{V^2})$.

(60) Approximately 35.287911086568917141298661140 2.

(61c) $\dfrac{\frac{\ell\sqrt{2}}{v_0}}{\frac{2g\ell}{v_0^2} - 1}$.

(63a) Some variation on $\ell = \dfrac{V^2}{g\cos\theta}(\sin 2\varphi - 2\cos^2\varphi\tan\theta)$.

(63b) $\frac{\pi}{4} - \frac{\theta}{2}$.

(68a) Not necessarily in order, 0.060 Hz, 0.38 rad/sec, 1.7 m/s^2, 17 sec.

(69a) 3.0×10^4 m/s.

(69c) 6.0×10^{-3} m/s^2.

(70) $\dfrac{T_e T_v}{T_e - T_v} = 1.5988209$ yr.

(71a) 3.0 m/s^2, 0.25 rad/s^2, 170 m/s^2.

(71b) 4.5.

(71c) 3.6 sec.

(71d) 1.05.

(72b) $\sqrt[3]{12\pi/A}$.

(72c) $\rho\sqrt[3]{18\pi^2 A}$.

(72d) $\rho(18\pi^2 A)^{\frac{2}{3}}$, $\quad \rho\sqrt[3]{12\pi A^2}$.

(73b) $v_x = \omega\rho(1 + \cos\omega t) \quad v_y = -\omega\rho\sin\omega t$
$a_x = -\omega^2\rho\sin\omega t \qquad a_y = -\omega^2\rho\cos\omega t$

(73(c)i) Velocity $2v$ forward, acceleration $\omega^2\rho$ down.

(73(c)ii) Velocity zero, acceleration $\omega^2\rho$ up.

(73(c)iii) Velocity $v\sqrt{2}$ $\searrow$, acceleration $\omega^2\rho$ backward.

(73(c)iv) Velocity $v\sqrt{2}$ $\nearrow$, acceleration $\omega^2\rho$ forward.

(78b) Spiral of Archimedes.

(78(c)i) $\omega A(\widehat{\mathbf{r}} + \omega t\widehat{\boldsymbol{\theta}})$.

(78(c)ii) $-\omega^3 At\,\widehat{\mathbf{r}} + 2\omega^2 A\,\widehat{\boldsymbol{\theta}}$.

(79(c)i) $2\beta At\,\widehat{\boldsymbol{\theta}}$.

(79(c)ii) $-4\beta^2 At^2\,\widehat{\mathbf{r}} + 2\beta A\,\widehat{\boldsymbol{\theta}}$.

(80a) $\dfrac{wv}{\ell\sin\theta + w\cos\theta}$.

(81a) 5 mph.

(81b) $\arctan\frac{4}{3}$.

(81d) $\frac{3}{2}$ mi.

(81e) $\arccos\frac{3}{4}$.

(81f) $2/\sqrt{7}$ hr.

(82) $\frac{\sqrt{7}}{2}v$ at $\sin^{-1}\sqrt{3/7}$ to either side of the ground-velocity vector.

(83a) $T = \dfrac{2\ell}{v}$.

(83b) $\dfrac{T}{1 - v_w^2/v^2}$.

(83c) $\dfrac{T}{\sqrt{1 - v_w^2/v^2}}$.

Chapter 4

Dynamics

4.1 Newton's Laws

Mechanics, the physics of the motion of bodies, is governed by Newton's three laws.[1] Whereas kinematics merely describes motion — where an object is, how fast it is moving, how this speed is changing, etc. — , dynamics is concerned with the causes of motion. According to Newton's laws, it is force that gives rise to motion.

The Newtonian definition of force coincides with our naive, intuitive notion of a force as simply a push or pull. Since you always push or pull something in a particular direction, force is a vector quantity.

Newton's force laws are as follows:

I. A body moves at a constant velocity (that is, at a constant speed in an unvarying direction) as long as no net force acts on it.

   This first law, known as the law of inertia, is actually a special case of the second law (the case $\mathbf{F} = 0 \Rightarrow m\mathbf{a} = 0 \Rightarrow \mathbf{a} = 0 \Rightarrow \mathbf{v}$ is not changing), and you will therefore have little use for it. Perhaps Newton

[1] This is actually a huge lie. In truth, the fundamental principle underlying all of physics is *symmetry*. The goal of "string" and other unification theories is to determine *the* symmetry that governs the universe; once this symmetry is discovered, all the physics of the universe, including the types of matter that can exist and how they can interact — in short, everything that can be or happen — is fully determined. As shown in Appendix B, the symmetry of invariance under spacetime translations directly gives rise to conservation of energy and momentum, which are therefore fundamental physical quantities. Force turns out to be just the derivative of energy and is therefore neither fundamental nor particularly useful for understanding the universe; it is just a historical artifact, a bad habit of thought that refuses to go away. Though admittedly it remains useful in mundane, engineering applications.

was simply hedging his bets on the second law.[2]

An example of the law of inertia is what frequently happened in car accidents in the days before seat belts and air bags: if you were cruising along at 60 mph when your car hit a tree, there was no restraint to exert a force on you, with the result that while the car very quickly lost its velocity, you would retain yours — moving in the same forward direction at the same 60 mph until you encountered the windshield and dash board.

II. In words: the net force on a body equals the mass of the body times the body's acceleration. Algebraically: $\mathbf{F} = m\mathbf{a}$.

The $\mathbf{F}$ *in* $\mathbf{F} = m\mathbf{a}$ *stands for the net force.*[3] If multiple forces are acting on a body simultaneously, what determines the evolution of the body's motion is the *vector sum* of the individual forces.

The $\mathbf{a}$ in $\mathbf{F} = m\mathbf{a}$ is the body's acceleration, with the same definition $\mathbf{a} = d\mathbf{v}/dt$ (rate of change of velocity) that we have been using all along. It is very important to note that what the net force determines at any given moment is *not* the body's velocity, but how the body's velocity is *changing*: the immediate effect of a force is a *change* in the body's motion. Of course, a body's velocity is the cumulative result of changes to its motion caused by forces acting at *previous* times, but the velocity a body has at any particular instant is irrelevant to Newton's second law; the net force on the body determines only how that velocity is changing at that instant, that is, whether the body is speeding up, slowing down, or changing its direction of motion. When you toss a ball straight up, for example, the only force acting on the ball after it leaves your hand is (neglecting air resistance) the downward force of gravity, and this results in the familiar downward acceleration g throughout the motion — a force and acceleration that are independent of the ball's velocity. Although the force with which your hand originally propels the ball upward determines the ball's initial velocity as it leaves your hand and therefore such things as how high the ball will ultimately rise and how fast it will be moving at any given point along its trajectory,

[2] Actually, there are historical reasons why Newton may have wanted to plainly state this first law: in the rather silly Aristotelian view that dominated Western thought for many centuries, it was believed that bodies had "natural" motions and that in the absence of a propelling force they tended to come to rest. And one could also get into a discussion of how spatial distances and time intervals, which are crucial to ascertaining that a body's velocity is constant, are fundamentally determined to begin with.

[3] This is one of those things that you might consider burning this into the back of your skull. Overlooking this point is the cause of a great many errors.

while the ball is in the air the force of gravity concerns itself, so to speak, solely with how the ball's velocity is changing: while the ball is on the way up, the downward force of gravity causes it to slow down; when the ball is at rest at its apex, gravity causes it to start to move downward; and while the ball is on the way down, gravity causes it to speed up. At each point along the ball's trajectory, the force of gravity is dictating, not the ball's velocity, but how that velocity is changing.

The m is the inertial mass of the body — the familiar mass you measure with a scale in MKS units of kilograms. As you can see from $\mathbf{F} = m\mathbf{a}$, for a given $\mathbf{F}$, the larger m is, the smaller $\mathbf{a}$ will be and vice versa. That is, the larger the inertial mass m, the less effect the applied force has on the motion. Hence the term *inertial*, as in *inertia*: the inertial mass measures a body's resistance to being set in motion or to changing its motion.[4] You are of course all familiar with the concept of inertia from study hall.

This second law can be applied either to individual bodies or to systems (that is, collections of bodies considered as a single entity). It can be shown that for these so-called extended bodies, the point for which $\mathbf{F} = m\mathbf{a}$ gives the acceleration is a special point known as the center of mass, which you will learn about in Chapter 6. Until then, you should just be aware that the acceleration $\mathbf{a}$ is strictly speaking the acceleration of only this one special point in the body.

III. Bodies exert equal and opposite forces on each other.

This explains why, for example, you can end up breaking your bat when you try to smash someone's skull: as the bat exerts tremendous force on the skull, the skull necessarily exerts an equally tremendous force back on the bat.

While not quantitative, this third law is critical to understanding the mechanical behavior of bodies. But it really is as simple as it sounds: bodies exert equal and opposite forces on each other. Period. End of story. The error into which people frequently fall is simply to forget to apply the third law. So just don't forget.

Many books state the third law as "For every action there is an equal and opposite reaction." To state the third law in this way should be a criminal offense; the third law applies to forces, so why confuse the

[4] In addition to inertial mass, there is also gravitational mass, the mass on which gravity acts. It turns out that inertial and gravitational mass are identical, a coincidence established by general relativity, and about which we will say more later.

issue by referring to "actions" and "reactions"? What the *%#$@! are those supposed to be? [5]

The MKS units of force are Newtons (N):

$$[F] = [m]\,[a] = \mathrm{kg} \times \frac{\mathrm{m}}{\mathrm{s}^2}$$
$$1\,\mathrm{N} = 1\,\frac{\mathrm{kg\,m}}{\mathrm{s}^2}$$

The English pound (lb) is also a unit of force: $1\,\mathrm{lb} = 4.448\,\mathrm{N}$.

4.2 Special Forces

Constant velocity & vanishing net force. When an object moves at a constant velocity (including the special case of remaining at rest), the net force on the object must vanish ($\mathbf{F}_{net} = 0$): because the object's velocity is not changing, its acceleration is by definition zero, so that $\mathbf{F} = m\mathbf{a} = 0$.

Be careful to note, however, that when an object is only at rest *for an instant*, its acceleration and the net force on it need not vanish. Consider again the case of a ball thrown straight up: throughout its entire flight, it is experiencing the usual $9.8\,\mathrm{m/s^2}$ downward acceleration due to gravity — on the way up, on the way down, and in particular also at the highest point, where it is for an instant at rest.

[5] Actually, there are historical reasons for this terminology: *action* and *reaction* are fairly literal translations of the Latin with which Newton expressed the third law in his *Philosophiæ Naturalis Principia Mathematica*:

I. *Corpus omne perseverare in statu suo quiescendi vel movendi uniformiter in directum, nisi quatenus illud a viribus impressis cogitur statum suum mutare.*

II. *Mutationem motus proportionalem esse vi motrici impressæ, et fieri secundum lineam rectam qua vis illa imprimitur.*

III. *Actioni contrariam semper et æqualem esse reactionem: sive corporum duorum actiones in se mutuo semper esse æquales et in partes contrarias dirigi.*

A fairly literal paraphrasing of the above (if such an oxymoron is possible) would be

I. Every body persists in its state of rest or of motion uniform in direction, except to the extent that it is compelled to change its state by applied forces.

II. The change in the motion is proportional to the applied force, and is in a straight line along that applied force.

III. To an action there is always an opposite and equal reaction: or, the actions of two bodies on each other are always equal and aligned in opposite directions.

Weight. Near the Earth's surface[6] the downward acceleration due to gravity is very nearly constant and has the value g. Thus the weight of a body, which by definition is just the force exerted on it by gravity, is

$$F_{grav} = ma_{grav} = mg$$

Be careful to distinguish the mass m from the weight mg; weight is a force (in N or lbs), while a mass is ... well, a mass (in kg).[7]

Normal forces. Whenever a body and a surface are in contact with each other, the equal and opposite forces that they exert on each other can always be broken into components parallel and perpendicular to the surface. By convention, the parallel component is (for better or worse) called friction, while the perpendicular component is called the normal force (*normal* here in the mathematical sense of *perpendicular*).

An example involving normal forces and friction will be worked out on p.202.

Cords and Pulleys. A tension is simply another variety of force: the tension in a segment of rope or cable is the pull that each its ends exerts on whatever it is connected to. To keep things simple, we will almost always make believe that ropes, cords, cables, etc., are massless. For such idealized cords, the tension is constant throughout any length of the cord between bodies (that is, from one body to the next, or from body to pulley, or from pulley to pulley). To see how this comes about, consider a small segment somewhere along such a length of cord: since this segment is massless, the net force on it must vanish, even if it has a nonzero acceleration: $F_{net} = ma = 0$. Since the only forces acting on this segment are the pulls from the tensions on either side of it, these tensions must therefore cancel — that is, the tensions on either side of the segment must be equal. And since this is true for all of the segments that make up the length of cord, the tension must be the same throughout the length of cord.

Real pulleys have mass and at least some friction, in which case the tension in the cord will generally differ between the two sides of the pulley. We will not, however, be able to take this into account until we get to rotational motion. In the meantime, we will restrict ourselves to the simple idealized case of the massless, frictionless pulley, for which the pulley serves only to redirect a tension that is the same on both sides of the pulley.

[6] Or near the surface of any spherically symmetric mass, with the value of g appropriate to that mass. How the value of g is determined we will see to on p.209, where it will also become clear that *near the surface* means that your distance from the surface is small compared to the radius of the sphere of mass. And this is of course ignoring complications due to the Earth's rotation and local geology (see footnote 14 on p.135).

[7] For those of you who were curious, there is an English unit of mass. Believe it or not, it is the *slug*: 1 slug=14.59 kg.

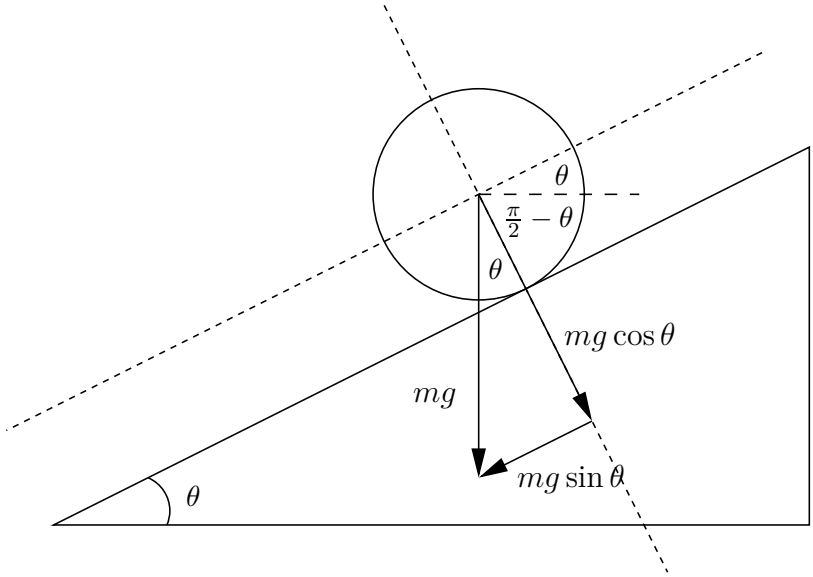


Figure 4.1: Components of mg on an Incline

An example involving a pulley and tension in a rope will be worked out on p.197.

Inclined planes. The pull of gravity on a body is, of course, always directly vertically downward. But when a body is on an incline, in general the motion will be along the incline, and in working out that motion we will therefore want to analyze our force vectors into components along tilted axes that run parallel and perpendicular to the incline, as shown in fig. (4.1). Memorization is almost always evil, but since you need to work with this kind of motion so frequently, it is good simply to remember that for a plane inclined at angle θ to the horizontal, the component of gravity down the plane is always $mg \sin\theta$ and the component perpendicular to the plane $mg \cos\theta$, as shown in fig. (4.1).

Bogus friction. Frictional forces are in reality very complex. For an object on a surface or two objects in contact, there is unfortunately no simple relation for friction that is even a decent approximation, and, if it were up to us, we would therefore not do very much with friction quantitatively. But it isn't up to us, so we will actually be doing a lot with a couple of utterly bogus "laws" of friction.

Since the origin of friction is the bonding between objects when they are in contact, it is, however, valid to distinguish between the cases of static friction, when there is no motion, and kinetic friction, when the objects are sliding on each other. Because stronger bonds can form when there is no motion, static friction is always at least as great as kinetic friction.

Now for the adventure in bogosity. For the case of an object on a surface, introductory texts usually give

$$F_{static\ friction} \leq \mu_s N$$
$$F_{kinetic\ friction} = \mu_k N$$

where N is the normal force between the body and surface and the μ are constants of proportionality (called the coefficients of friction) that differ from one body and surface to another. These "laws" of friction are both a gross oversimplification and a poor approximation, but you are subjected to them almost conspiratorially by introductory physics texts because they want you to have relations for friction simple enough for you to handle in calculations, even if those relations are total malarkey. It's all damnable lies. But, hey, we don't make the rules.

Anyway, the $\leq$ sign in the relation for static friction is to take into account that there will only be a static frictional force if there is some applied force to oppose. For a box at rest on a floor, for example, there will not be any frictional force unless you push on the box to try to move it, and then this frictional force will, up to its maximal possible value $\mu_s N$, adjust itself to cancel out the force you are applying and keep the box at rest. If you apply a force greater than $\mu_s N$, you will have overcome static friction; the box will begin to move, and kinetic friction will kick in.

Examples involving bogus frictional forces will be worked out on pp.197 and 202.

4.3 Force Diagrams

When trying to analyze and understand the dynamics of a system's motion, it is helpful to draw for each body in the system a force diagram[8] that shows the various forces acting on that body. This diagram will make it easier to apply $\mathbf{F} = m\mathbf{a}$ by helping you see the directions of the force vectors and how each of these vectors will contribute to the vector sum that constitutes the net force.

The force diagram for a body should include all (and only) the forces acting *directly* on that body. Since the more complicated and cluttered your diagram, the less clear and helpful it will be, you should for simplicity represent the body simply by a dot and the forces as arrows, labeled to make clear to which force each arrow corresponds.[9]

[8] The term free-body diagram is used by some books.

[9] At least for the time being. When we get to rotational motion, we will see that it matters just where on the body the various forces act. But that complication lies in the future.

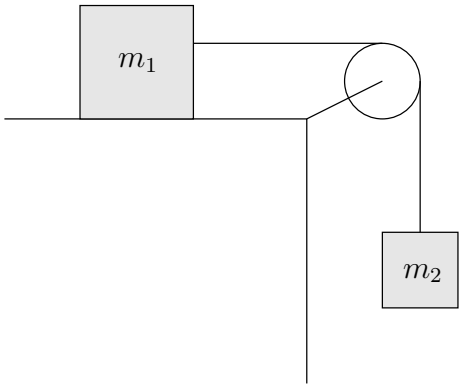


Figure 4.2: An Utterly Lame Arrangement of Masses

For example, suppose we have the utterly lame and uninspiring arrangement of dumb masses connected by a cord slung over a pulley shown in fig. (4.2). What is the world coming to? Anyway, for the case that the mass m_2 is descending, the corresponding force diagrams are shown in fig. (4.3). Not likely to get you an A in an art course, but ideal for analyzing forces and setting up $\mathbf{F} = m\mathbf{a}$.

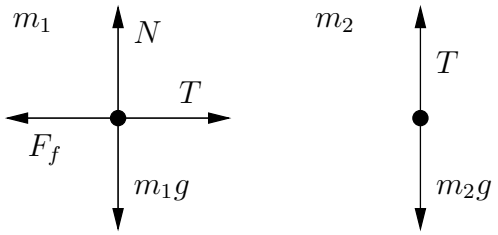


Figure 4.3: The Corresponding Force Diagrams

In the diagram for m_2, we have noted that there are only two forces acting directly on m_2: its own weight $m_2 g$ pulling straight down, and the tension T in the cord pulling straight up. Since both forces are vertical, there is only this one axis along which to apply $\mathbf{F} = m\mathbf{a}$, and if we take up to be the positive direction, we therefore have

$$T - m_2 g = m_2 a \tag{4.1}$$

In the diagram for m_1, we have noted that there are four forces acting directly on m_1: its own weight $m_1 g$ pulling straight down; the tension T in the cord pulling horizontally to the right; and a force, exerted on m_1 by the surface with which it is in contact, that we have split into a perpendicular supporting (normal) force N and a parallel frictional force F_f. We know that the frictional force F_f will be to the left because it was given that m_2

was descending: because of the cord connecting the two masses, downward motion of m_2 entails m_1 moving to the right, and friction, to oppose this motion, will therefore be toward the left.

The force diagram for m_1 is thus two-dimensional. Since the motion of m_1 is along the horizontal surface, we want to analyze our forces and set up $\mathbf{F} = m\mathbf{a}$ along the horizontal and vertical directions. In the horizontal direction, the net force will be either $T - F_f$ or $F_f - T$, depending on whether we take to the right or to the left to be the positive direction. Usually the choice is arbitrary. Here, however, the motions of two masses are coupled by the cord, so that m_1 and m_2 will move the same distance and share the same speed and acceleration. Since we already chose up to be the positive direction for m_2, if we want to use the same acceleration a in our horizontal equation for m_1, we must, to be consistent, take to the left to be the positive direction for m_1. Then $\mathbf{F} = m\mathbf{a}$ along the horizontal direction becomes

$$F_f - T = m_1 a \tag{4.2}$$

And if we take up to be positive, then along the vertical direction $\mathbf{F} = m\mathbf{a}$ becomes [10]

$$N - m_1 g = m_1 a_{vert} \tag{4.3}$$

Our force diagrams have now served their purpose by enabling us to set up $F = ma$ and generate eqq. (4.1)-(4.3). The unknowns in these equations are the accelerations a and a_{vert}, the normal force N, the tension T, and the frictional force F_f — five unknowns, for which we have only three equations. We therefore need two other pieces of information. First, we note that physically we do not expect there to be any vertical acceleration of m_1: m_1 is not going to sink down into the surface or leap up off of it. Therefore $a_{vert} = 0$ and eq. (4.3) yields

$$N = m_1 g$$

That is, the supporting normal force exerted by the surface on m_1 exactly balances the weight of m_1.

Second, since m_1 is moving over the surface, the frictional force is kinetic, and so (in the perversion of reality presented by introductory textbooks) $F_f = \mu_k N = \mu_k m_1 g$. Using this result for the frictional force in eqq. (4.1)

[10] There is no issue of consistency for the vertical direction because the vertical motion of m_2 is coupled to the *horizontal* motion of m_1, not to the vertical motion of m_1. Had we set everything up without taking into account that the motions of m_1 and m_2 are coupled by the cord, we would have labeled three distinct accelerations: a for the vertical acceleration of m_2 and a_{horz} and a_{vert} for the horizontal and vertical accelerations of m_1. The coupling of the horizontal motion of m_1 with the vertical motion of m_2 has allowed us, since we were consistent about directions, to set $a_{horz} = a$, but a_{vert} remains independent.

and (4.2) then gives us, after a little straightforward algebra,

$$a = -\frac{m_2 - \mu_k m_1}{m_1 + m_2} g \tag{4.4a}$$

$$T = (1 + \mu_k)\frac{m_1 m_2}{m_1 + m_2} g \tag{4.4b}$$

We see from this solution that T is always positive, as must be the case for a tension.[11] The interpretation of our solution for a, however, falls into three separate cases:

- For $m_2 > \mu_k m_1$, a is negative, which, since we made up the positive direction, corresponds to m_2 accelerating downward. Since it was given that m_2 was moving downward, this means that m_2 is speeding up as it falls.
- For $m_2 = \mu_k m_1$, $a = 0$. In this case, the speed at which m_1 and m_2 are moving remains constant.[12]
- For $m_2 < \mu_k m_1$, a is positive, which corresponds to m_2 accelerating upward. Since it was given that m_2 was moving downward, this means that m_2 is slowing down as it falls. Eventually the masses will come to rest, at which point static friction will kick in. Since kinetic friction was strong enough to bring the masses to rest and static friction is always at least as great as kinetic friction, static friction will be enough to keep the masses at rest.

In other words, what happens depends on the values of the parameters m_1, m_2, and μ_k. As you can see from the numerator of eq. (4.4a), m_2 tends to make a negative, which corresponds physically to the weight of m_2 being what pulls the masses along. μ_k and m_1 tend to make a positive, which corresponds physically to friction tending to hinder the motion. The amount of friction depends directly on μ_k, which gives a measure of the degree of adhesion between m_1 and the surface, and indirectly on m_1: the greater m_1, the greater the force of contact $N = m_1 g$ between m_1 and the surface and hence the greater the frictional force $F_f = \mu_k N$. The m_1 and m_2 in the denominator of eq. (4.4a) ultimately came from the ma side of the force relations and therefore correspond to inertial effects: that m_1 is added to

[11] A negative tension would correspond to the cord *pushing* rather than pulling on the masses, which cords can't do. Rods and the like can, however, push as well as pull the objects to which they are connected, and in this sense a negative tension constitutes a stress. Stresses are of huge importance in engineering, but they are beyond the scope of this text and we will not be dealing with them any further.

[12] This is an example of what is known technically as a critical case. A critical case is a special case that separates domains of differing behavior. In this particular example, the case $m_2 = \mu_k m_1$ separates solutions with $a > 0$ from those with $a < 0$.

m_2 in this denominator indicates that m_1's inertia is slowing down (reducing the magnitude) of the acceleration, corresponding to m_2's having to drag m_1 along with it.

We could take a detailed, quantitative look at various limits of the parameters, among them the case $m_2 \to \infty$. In this limit, eqq. (4.4a) and (4.4b) become

$$\begin{aligned} a &= -\frac{m_2 - \mu_k m_1}{m_1 + m_2} g \\ &\to -\frac{m_2}{m_2} g = -g \\ T &= (1 + \mu_k) \frac{m_1 m_2}{m_1 + m_2} g \\ &\to (1 + \mu_k) \frac{m_1 m_2}{m_2} g = (1 + \mu_k) m_1 g \end{aligned}$$

This limit of the result for acceleration makes perfect sense physically: as the mass hanging over the side becomes infinite, it approaches free-fall — by comparison to the weight of m_2, the hindrance presented by the inertia of m_1 and by the friction between m_1 and the surface are negligible. The limit of the result for the tension also makes sense: the tension goes to the sum of $m_1 g$ (the force necessary to give m_1 an acceleration of g) and $\mu_k m_1 g$ (the force necessary to counter the friction between m_1 and the surface).

Finally, a couple of other very important points raised by this example:

- Note that the weight of m_2 is not included in m_1's force diagram: the weight of m_2 is a force exerted on m_2, not on m_1, and although ultimately it is the weight of m_2 that pulls m_1 along, m_1 does not know anything directly about the weight of m_2; all m_1 knows is that it feels the tug of the tension in the cord.

- In this example we are not concerned with the forces acting *on* the surface, but since, as we have just found, the surface exerts an upward force $N = m_1 g$ on the mass m_1, the third law tells us that m_1 must necessarily exert an equal and opposite force on the surface. That is, m_1 is exerting on the surface a downward force equal in magnitude to the weight of m_1. But just because the force exerted on the surface is equal in magnitude to the weight of m_1 does *not* mean that the weight of m_1 *is* the force acting on the surface: the weight of m_1 is a force exerted gravitationally by the Earth on m_1, not on the surface. The surface does not know anything directly about the weight of m_1; all the surface knows is that it feels a downward *normal force* exerted on it by m_1. *The weight of m_1 does not act on the surface.* While in this particular example the normal force felt by the surface equals the

weight of m_1, under other circumstances that equality would not even hold: the result $N = m_1 g$ came about because $a_{vert} = 0$, and a_{vert} would not be zero were this arrangement of the two masses inside, for example, an accelerating elevator.

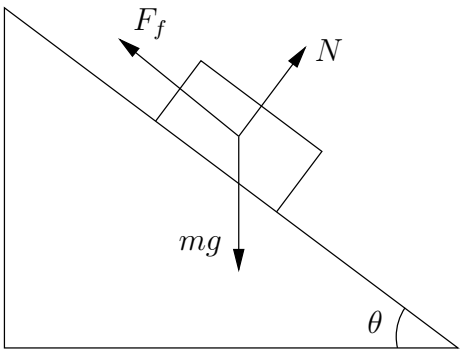


Figure 4.4: A Stupid Block on a Stupid Incline

As another example, consider a stupid block held at rest on a stupid incline by friction, as shown in fig. (4.4). Since the block would otherwise slide down the incline, we know that friction, to prevent this sliding, must be up the incline. Although we could analyze the forces of this two-dimensional situation along any two linearly independent axes, let us analyze them along tilted axes that run parallel and perpendicular to the incline. If we make down the incline the positive direction for the axis parallel to the incline, then the components of the forces in the parallel direction are (recall fig. (4.1))

$$\begin{aligned} N_{\parallel} &= 0 \\ F_{f\parallel} &= -F_f \\ (mg)_{\parallel} &= mg \sin\theta \end{aligned}$$

so that $F = ma$ along this direction becomes

$$mg \sin\theta - F_f = ma_{\parallel} \tag{4.5}$$

And along the perpendicular direction, if we take out of the incline as the positive direction, the components are

$$\begin{aligned} N_{\perp} &= N \\ F_{f\perp} &= 0 \\ (mg)_{\perp} &= -mg \cos\theta \end{aligned}$$

so that $F = ma$ along this direction becomes

$$N - mg \cos\theta = ma_{\perp} \tag{4.6}$$

It was given that the block was being held at rest by friction, but even if there were motion, that motion would be entirely parallel to the incline; there would be no motion perpendicular to the incline. Therefore $a_\perp = 0$, so that eq. (4.6) yields [13]

$$N = mg\cos\theta \tag{4.7}$$

In this particular example, since it was given that the block was being held at rest by friction, we also have $a_\parallel = 0$. Eq. (4.5) thus reduces to

$$F_f = mg\sin\theta \tag{4.8}$$

Eq. (4.8) is telling us that if the block is to be held at rest, friction must balance the component of the weight pulling the block down the incline. Without more information, this is as far as we can go in our analysis. Suppose, however, that we also know that the block is just on the verge of slipping — that if we tilt the incline to any steeper angle, it would begin to slip. Then the static friction F_f between the block and the incline would have reached its maximal possible value, so that $F_f = \mu_s N$. Using this relation for friction and our result (4.7) for the normal force in eq. (4.8), we have

$$\mu_s mg\cos\theta = mg\sin\theta$$

which yields

$$\tan\theta = \mu_s$$

The steepest angle at which friction can hold the block at rest is therefore determined solely by the coefficient of bogus friction μ_s and does not depend on the mass of the block or on g — blocks of all masses will be on the verge of slipping at the same angle θ, and this will be the case on the Moon as well (even though g on the Moon is only about $\frac{1}{6}$th what it is on Earth).

Note however that we are able to set $F_f = \mu_s N$ only when we know that the block is on the verge of slipping; if the block is not on the verge of slipping, less friction than the maximal value $\mu_s N$ is required to keep the block at rest and therefore it would be wrong to set $F_f = \mu_s N$. Be forewarned: while the force of bogus kinetic friction is always $F_f = \mu_k N$, reflexively setting the force of bogus static friction $F_f = \mu_s N$ is a very common error.

[13] A very common error is to assume that the result (4.7) always holds. While $N = mg\cos\theta$ in this and in fact in most cases, this would not be true if there were another force in the problem with a perpendicular component — as would be the case if, as in problem #17, you were applying a horizontal push or pull to the block: your horizontally applied force would have a perpendicular component and would therefore enter into eq. (4.6), changing the result we would obtain for N. Nor would eq. (4.7) hold when there is somehow a contribution to the block's acceleration that leaves $a_\perp = 0$ nonzero, as would be the case if, as in problem #18, the block and incline were in an accelerating elevator.

4.4 Circular Motion

According to eqq. (3.32) and (3.36), a body moving in a circle must, simply by virtue of its circular motion, experience the centripetal acceleration

$$a_{cent} = \omega^2 r \quad \text{or, equivalently,} \quad \frac{v^2}{r}$$

in toward the center of the circle. So if we know that a body is moving in a circle, we automatically know the acceleration a on the ma side of $F = ma$:

$$\begin{aligned} F &= ma_{cent} \\ &= m\omega^2 r \quad \text{or, equivalently,} \quad \frac{mv^2}{r} \end{aligned} \tag{4.9}$$

Eq. (4.9) is telling us that a body that is moving in a circle must be experiencing a *net radial force in toward the center* of the circle and that this net radial force must equal ma_{cent}. The question is simply what physical forces give rise to this force. In other words, when we set up $F = ma$ along the radial direction for circular motion, we want to add up the radial components of the forces acting on the body, counting those toward the center as positive and those away from the center as negative, and then set this net radially inward force equal to ma_{cent}.

For example, suppose on a roller coaster you are traveling at speed v around a classic circular loop-the-loop of radius r. When you go around the bottom of the loop, your force diagram is as shown in fig. (4.5): your weight mg vertically downward and a vertically upward normal force N exerted on you by the seat. Both of these forces are radial, with N being toward the center of the loop and mg away from it. Thus $F = ma$ becomes

$$N - mg = \frac{mv^2}{r}$$

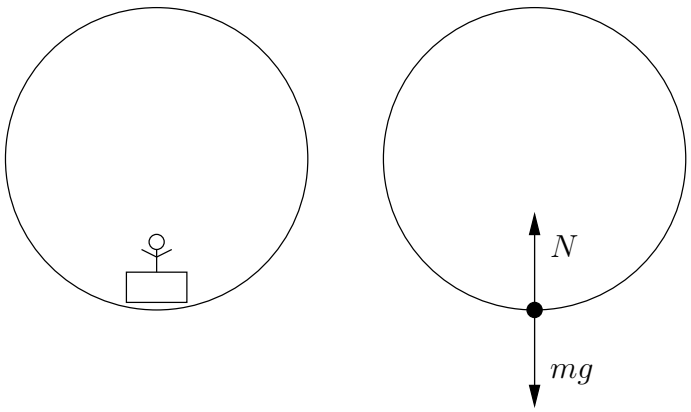


Figure 4.5: Bottom of a Loop-the-Loop

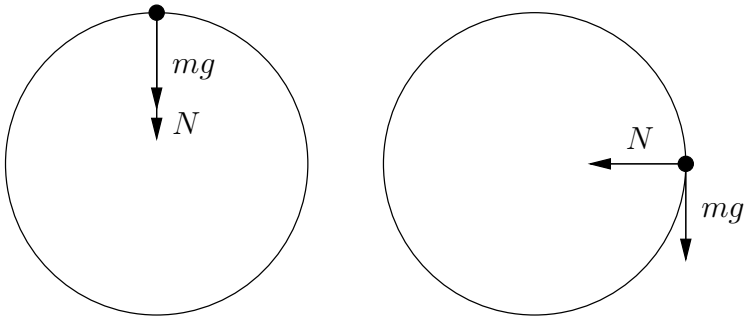


Figure 4.6: Top and Midpoint of a Loop-the-Loop

When you are at the top or halfway up, these same two forces are acting on you, and your force diagrams are as shown in fig. (4.6). Thus at the top, where both N and mg are in toward the center of the circle,

$$N + mg = \frac{mv^2}{r}$$

And at halfway up, N is in toward the center, but mg is tangential — it has no radial component and therefore does not contribute to centripetal effects. Thus

$$N = \frac{mv^2}{r}$$

It is conventional to refer to ma_{cent} as the centripetal force F_{cent}, but you need to bear in mind that *centripetal force is not a physical force* like a tension or a weight; it's just an expression for what the ma side of $F = ma$ works out to in the case of circular motion. You should therefore not draw an arrow labeled F_{cent} in your force diagrams, since there really isn't any such force.

Which brings us to centrifugal force and the issue of physical reality versus subjective perceptions. Consider a car turning in a circular arc. If we know the radius r of the turn and the speed v at which the car is taking it, then we know that the car's acceleration is $a_{cent} = v^2/r$ and that, according to $F = ma$, the net radially inward force on the car must come out to $F = mv^2/r$, where m is the car's mass. All of this we know simply because it is given that the car is moving in a circle; what physical forces give rise to this centripetal force needed to keep the car moving in a circle is an entirely separate question. In this case, the physical force turns out to be friction between the tires and the road: were there no friction, as on wet ice, the car would not move in circle; without any force to cause a change in its velocity, the car would simply move along a straight line at a constant speed. So the reality is this: when a car is making a turn, friction between the tires and the road is pushing the car toward the center of the turn.

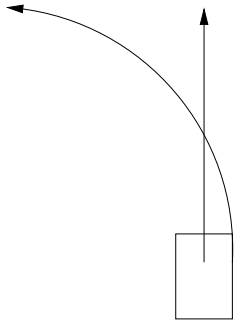


Figure 4.7: The Car Door and the Greased Butt

If you are in a car that is turning in a circular arc, you, too, are moving in a circle and are therefore experiencing a centripetal force $F = mv^2/r$ in toward the center of the turn. The physical force that supplies this needed centripetal force is the friction between the seat of your pants and the seat of the car: without this friction, you cannot make the turn with the car.

Subjectively, however, you do not feel yourself pulled in toward the center of the turn. What you feel is in fact just the opposite: a centrifugal force throwing you toward the outside of the turn. But this subjective perception is simply wrong: physically, there is no such thing as a centrifugal force;[14] you are simply experiencing inertial effects and attributing them to the action of a nonexistent force. If you were to grease your butt and then go around a turn in a car, you would slide across the seat and get slammed into the door, and it would feel like this happened because of a force throwing you toward the outside of the turn. The reality is otherwise: friction between the tires and the road is supplying the centripetal force needed to move the car in the circular arc of the turn, but when your butt is greased you are not experiencing the centripetal force needed to make you move in that arc. The result is that the car makes the turn, while you, like any body on which no force acts, continue moving in a straight line with your original velocity. When the circular path followed by the car door and the straight path you are following intersect, the collision occurs. Fig. (4.7) shows this as seen from above: the rectangle is the outline of the car at the instant it begins the turn, the arc is the path followed by the car door in the circular motion that it shares with the rest of the car, and the straight line is the path followed by you, who simply move at constant velocity and do not share in that circular motion. Though it feels like you are being thrown outward and smashing into the door, in fact it is the door that is veering around and smashing into you.

[14]Such forces are technically termed fictitious and are artifacts of your perceiving things from the perspective of a noninertial reference frame (that is, an accelerating reference frame) in which $F = ma$ does not hold.

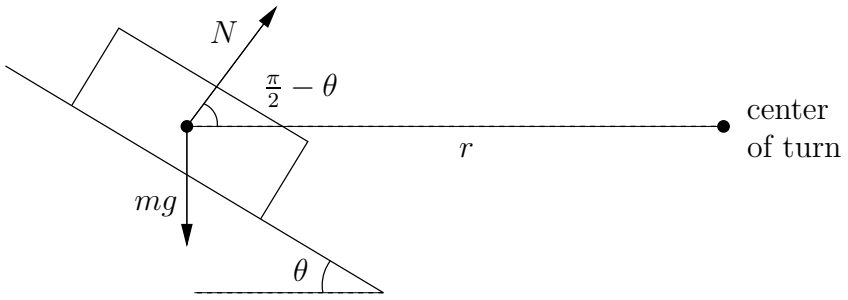


Figure 4.8: Ideal Road Banking

4.4.1 Road Banking

Fig. (4.8) shows, in cross section, a car going around a banked turn. Or at least that's what it's supposed to show. The line tilted up at angle θ represents a slice through the road surface and the rectangle is the car, which at the instant illustrated is coming out of the page toward you.

The purpose of banking the road is to help cars make the turn without the need for as much friction between the tires and the road. The road banking is said to be ideal when there is no need for any friction at all, so that the car could make the turn even if the road surface were covered with wet ice. In this ideal case, the only forces acting on the car are its weight mg and the normal N exerted on the car by the road.

Usually when a body is on an incline we work with axes tilted along the parallel and perpendicular directions, because any motion is along the incline. Here, however, we have motion in a horizontal circle, so we want to set up $F = ma$ along the horizontal and vertical directions.

In fig. (4.8) we have labeled the angle between N and the horizontal as $\frac{\pi}{2} - \theta$. The quick and dirty way to see this is as follows: we know that the angle must be either θ or $\frac{\pi}{2} - \theta$, and to see which, just imagine that the angle θ at which the road is tilted is really small. In this case, N will be nearly vertical, so that the angle between N and the horizontal will be nearly $\frac{\pi}{2}$. So we want $\frac{\pi}{2} - \theta$. The horizontal and vertical components of N are thus

$$N_{horz} = N\cos(\frac{\pi}{2} - \theta) = N\sin\theta$$
$$N_{vert} = N\sin(\frac{\pi}{2} - \theta) = N\cos\theta$$

If we are moving in a horizontal circle, there is no vertical motion, so the acceleration in the vertical direction must be zero. With up as the positive direction, $F = ma$ along the vertical axis becomes

$$N\cos\theta - mg = 0 \tag{4.10}$$

The horizontal component of N is toward the center of the circle, so $F = ma$ along the horizontal axis becomes

$$N \sin\theta = \frac{mv^2}{r} \tag{4.11}$$

Eliminating N to get a relation between the banking angle θ and the ideal speed v at which to take the turn, we arrive at

$$\tan\theta = \frac{v^2}{gr}$$

A couple of observations about this result: First, there is only one speed for which the banking is ideal. If a car's speed differs from this ideal speed, it will need friction to make the turn — the farther its speed from the ideal, the more friction will be needed. Second, the relation between θ and v is independent of the mass m of the car. This means that the mass of the vehicle need not be taken into account when engineering the banking for a turn: the same banking that is correct for a Yugo will be correct for a Mack truck.

4.5 Newton's Law of Gravity & Orbits

Gravitational dynamics are governed by Newton's law of gravity:[15]

$$F_{grav} = \frac{Gm_1m_2}{r^2} \tag{4.12}$$

Eq. (4.12) gives the magnitude of the equal and opposite gravitational force that two masses m_1 and m_2 exert on each other. r is the distance between the two masses, and $G = 6.6742\times10^{-11}\ \mathrm{Nm^2/kg^2}$, the constant of proportionality, is the universal gravitational constant.[16] This gravitational force is always an attraction: the two masses m_1 and m_2 always pull on each other.[17] Because this force is proportional to $1/r^2$, it is called an inverse-square law.

Most strictly, eq. (4.12) applies only to point masses, that is, masses concentrated at single points. Extended bodies, can, however, be regarded as

[15] This is yet another lie; gravitational dynamics are governed by general relativity (though general relativity still needs to be quantized — something no one as yet knows how to do). But as long as the masses involved are not too dense and the distance r is not too small compared to the masses, Newton's law of gravity is accurate to a very good approximation. In fact, Newton came up with this law of gravity by asking what gravitational force law would account for the motion of the planets, as we will see on p.211.

[16] So called because it is the same for any masses m_1 and m_2, whereas g is $9.8\ \mathrm{m/s^2}$ only for the Earth and has different values for different astronomical bodies.

[17] As opposed to the electric force, which can be either an attraction or a repulsion.

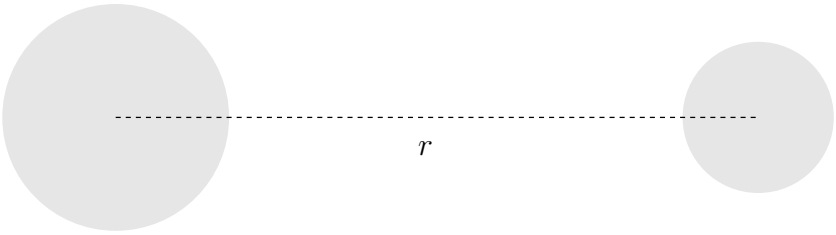


Figure 4.9: r in $\frac{Gm_1m_2}{r^2}$ for Spheres of Mass

sets of point masses, and, though it is far from obvious, it turns out that eq. (4.12) holds exactly as written also for spherically symmetric distributions of mass.[18] This means that to a good approximation, eq. (4.12) can be applied to the Sun, planets, and moons of our solar system. When dealing with spheres of mass, r is measured between the centers of the spheres, as shown in fig. (4.9).

$F_{grav} = mg$ is a special case of eq. (4.12): when you are near the surface of the Earth,[19] the distance r between the mass m_{obj} of an object and the center of the Earth is to a very good approximation simply the Earth's radius, so that

$$F_{grav} = \frac{Gm_{earth}m_{obj}}{r^2} \approx \frac{Gm_{earth}m_{obj}}{r_{earth}^2} = m_{obj}\left(\frac{Gm_{earth}}{r_{earth}^2}\right)$$

which, if we define

$$g = \frac{Gm_{earth}}{r_{earth}^2}$$

becomes

$$F_{grav} = m_{obj}g$$

If you plug in the values of G, m_{earth}, and r_{earth}, you will find that you do in fact get the familiar $9.8\,\mathrm{m/s^2}$.

Using Newton's law of gravity (and quite a bit more math than we have at our disposal), it is possible to prove all three of Kepler's laws analytically: [20]

[18] We could prove this by doing an integration over a spherical shell of uniform mass density, and this is in fact what Newton himself did — though it took him some twenty years because he had to invent calculus in order to do the integration. When we get to electromagnetism, however, Gauss's theorem and law will make this property of the gravitational force much easier to prove, and we will therefore simply accept the property for now and postpone its proof until p.784.

[19] This same prescription could, of course, be used to get a result for g for any spherically symmetric mass, not just the Earth.

[20] Actually, you will prove the third law, at least for circular orbits, in problem #45, and general proofs of all of Kepler's laws are given in Appendix C.

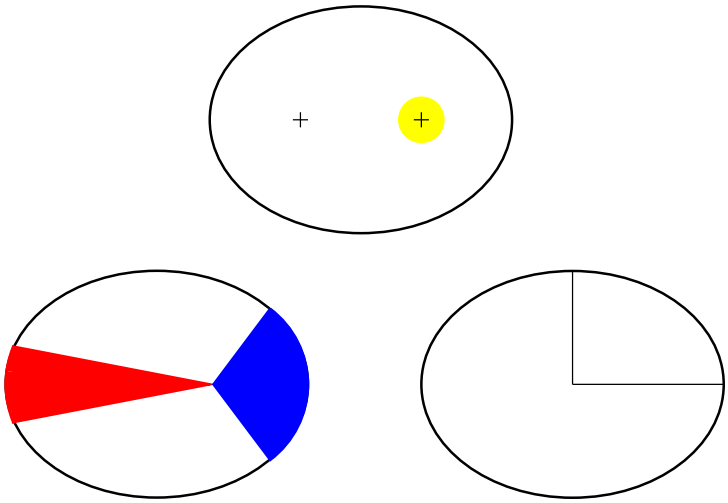


Figure 4.10: Kepler's Laws

I. The orbits of the planets are ellipses, with the Sun at one focus.[21]

II. As each planet orbits around the Sun, it sweeps out equal areas in equal times.

III. The square of the period of a planet's orbit is proportional to the cube of the semimajor axis of the ellipse.[22]

Fig. (4.10) is the kind of illustration of Kepler's laws that a third-grader might draw, except that the yellow ball that represents the Sun is lacking a smiley face. Anyway, the top of the figure shows the two foci of an elliptical orbit, the lower right the semiminor and semimajor axes. The lower left of the figure illustrates Kepler's second law: over the course of a month or some other given interval of time, the planet will (as shown by the blue wedge in fig. (4.10)) sweep out a sector of greater arc when it is closer to the Sun (and therefore, having been pulled in by the Sun's gravity, moving faster) than when (as shown by the red wedge in fig. (4.10)) it is farther away from the Sun and therefore moving more slowly. A detailed calculation would show that the areas of such sectors are, however, always exactly equal, with the longer arc of the stubbier sector exactly compensating for its shorter "radius."

We will not do anything further with Kepler's laws, but you should be aware of them; historically, they are, together with Newton's law of gravity, regarded as one of the principal early examples of the scientific method: a

[21] The question is often asked, What's at the other focus? Answer: Nothing. The other focus is just an abstract geometric point; physically there doesn't have to be anything at all there. Sometimes life is disappointing. You'll just have to suck it up.

[22] Since it is "squashed," an ellipse has two "radii": the shorter one is called the semiminor axis and the longer one the semimajor axis.

theory deriving from observed laws and in turn being tested through empirically verifiable predictions.

A special case of an ellipse is a circle, and in fact the orbits of most of the planets are very close to being circular.[23] In this case, gravity is the physical force supplying the centripetal force needed for the circular motion: if m is the mass of the "planet" orbiting at speed v around a "Sun" of mass M,[24]

$$\frac{GMm}{r^2} = \frac{mv^2}{r} \tag{4.13}$$

Eq. (4.13) relates the orbital speed v of the satellite to the radius r of the orbit: you can't have just any old circular orbit; for a given radius of orbit, there is only one special speed that will work.

If you are dealing with a circular orbit, you should be able to set eq. (4.13) up and work out what you need from it — rather than by memorizing or looking up a lot of "pre-fabricated" formulas that have very limited application. If you use eq. (4.13) in conjunction with the various relations that apply generally to circular motion, you can get any result you need. If, for example, you want to relate the radius of the orbit, not to its speed, but to its period, you can use $v = 2\pi r/T$ in eq. (4.13) to eliminate v and obtain an equation in terms of T and r. Etc., etc. Yada yada. Blah, blah, blah.

It was in fact through this sort of reasoning that Newton came up with his inverse-square law of gravity: Kepler's laws of planetary motion were already known, and Newton, having, as we saw in §3.3.3, already derived a result for the centripetal acceleration a body would experience in a gravitational circular orbit, was therefore able to start from

$$F_{grav} = ma_{cent} = \frac{mv^2}{r}$$

Using this relation in conjunction with $v = 2\pi r/T$ yielded

$$F_{grav} = \frac{m}{r}\left(\frac{2\pi r}{T}\right)^2 = \frac{4\pi^2 mr}{T^2}$$

[23] Only systems where all the planets have very nearly circular orbits have stable enough conditions for life, but this property actually makes our solar system unusual: orbits of planets are much more likely to be markedly eccentric, as we are finding is indeed the case for most planets outside of our own solar system. On the other hand, there are megagigabazillions of solar systems out there, so we certainly shouldn't flatter ourselves by thinking that we are in any way unique.

[24] We use quotes because these are relative terms: in the context of a satellite orbiting the Earth, the Earth is M and the satellite is m. Also, this is another lie. Actually, the two masses m and M orbit around each other, as is seen with binary stars. But if, as is the case with the planets of our solar system and the Sun and with Earth satellites and the Earth, M is far greater than m, then M will be nearly stationary. Eventually, when we have rotational dynamics under our belt, you will deal with a binary orbit in problem #43 of Chapter 7.

To pull a law of gravity out of this, we want to express F_{grav} as a function solely of r, not of r and T, and to accomplish this Newton made use of Kepler's third law, which for a circular orbit states that $r^3 \propto T^2$ — that is, that $T^2 = \text{const} \times r^3$ for some constant of proportionality. Thus

$$F_{grav} = \frac{4\pi^2 mr}{\text{const} \times r^3} = \frac{4\pi^2 m}{\text{const}} \frac{1}{r^2}$$

which establishes that the gravitational force does indeed go as $1/r^2$.

Note that when you are in a gravitational orbit, you are *not* weightless: your weight, in the sense of the gravitational force on you, is given by $F_{grav} = Gm_1m_2/r^2 \neq 0$. You are, however, falling freely under the influence of this gravitational force, with the result that, as will be explained in the next section, you *feel* weightless, just as you would in a freely falling elevator on Earth. These two very important points bear reiterating: contrary to common misconceptions, when you are in orbit, you are *not* weightless, but you *are* in free-fall.[25]

4.6 Perceived Weight

The gravitational pull that you and the Earth exert on each other is given by Newton's law of gravity,

$$F_{grav} = \frac{Gm_1m_2}{r^2}$$

If you are near the surface of the Earth, this gravitational force can also be expressed to a good approximation as $F_{grav} = mg$. In any case, your weight is the force that gravity exerts on you, and that force does *not* depend on what motion you happen to be undergoing.

Suppose, for example, that you are in an elevator, as shown in fig. (4.11): your force diagram will consist of your weight mg and the upward normal force N exerted on you by the floor of the elevator. With up as the positive direction, $F = ma$ thus becomes

$$N - mg = ma \tag{4.14}$$

In this relation, your mass m is constant and g is constant; the quantities that change depending on your motion are a and N. If you are moving at

[25]The objection is often raised, If you are in free-fall while in orbit, then why don't you fall in toward the Earth? Remember that the acceleration to which the gravitational force gives rise is the *centripetal* acceleration required for your circular motion and therefore corresponds, not to any motion in the radial direction, but to the continual change in the direction of your tangential velocity.

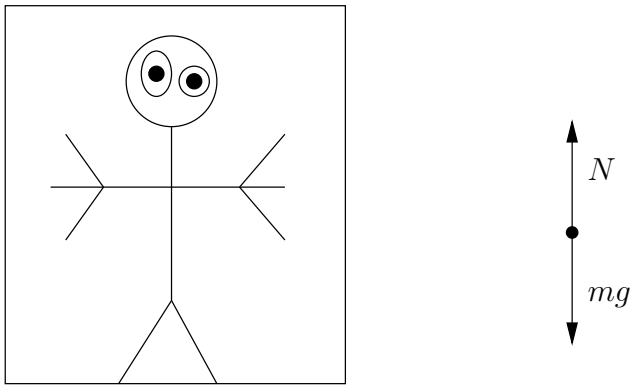


Figure 4.11: Life in an Elevator.

a constant speed, whether up or down, $a = 0$, the same as it would be if you were remaining at rest, and when $a = 0$ eq. (4.14) yields $N = mg$. If, on the other hand, the cable has snapped and you are in free-fall, you are accelerating at a full g downward, so that $a = -g$ and eq. (4.14) yields $N = 0$. That is, when you are in free-fall, the force of contact between you and the floor drops to zero.

In every case, your weight, in the sense of the gravitational force on you, is mg and that's that. Your *perceived* weight, however, is a different matter. Subjectively, you gauge how heavy you are by how hard you feel yourself pressed into the floor or the seat. This force of contact is in fact just the normal force N. When the elevator is sitting at rest, $N = mg$, corresponding to your feeling that you are your "normal" weight mg.[26] When the elevator is in free-fall, $N = 0$, corresponding to your feeling weightless. In all cases, it is the normal force N that determines how heavy you feel.[27]

As pointed out in the previous section, when you are in a gravitational orbit, your weight, in the sense of the gravitational force on you, is $F_{grav} = Gm_1m_2/r^2 \neq 0$: you are *not* weightless in orbit. You are, however, falling freely under the influence of this gravitational force, with the result that you *feel* weightless, just as you would in a freely falling elevator on Earth. What are termed zero-gravity experiments on the Space Shuttle and International Space Station would be more accurately termed free-fall experiments.[28]

[26] This collision between the mathematical and everyday uses of the term *normal* is unfortunate, but you'll just have to deal with it. Life is complicated sometimes.

[27] This is of course a lie — once again life is not so simple — , but as lies go it's not a really big, bald-faced one, and the "exceptional" cases can be dealt with by common sense. When, for example, you are hanging at rest from a rope, it is the upward pull of the tension T that equals your weight mg and that gives rise to the perception that you are your normal weight.

[28] Although, according to the principle of equivalence in general relativity (§10.9), the physical behavior of objects in free-fall is equivalent to their behavior in the absence of gravity, at least for small regions of spacetime.

4.7 Semi- & Almost Nonbogus Friction

For certain cases, such as a boat gliding through water, the frictional force is to a reasonably good approximation proportional to the velocity of the body. This is still semibogus, but it's not too far from reality. In the case of a boat, if this friction is the only horizontal force acting on the body, $F = ma$ becomes

$$-\beta v = ma \tag{4.15}$$

where we have written β for the (positive) constant of proportionality and have included the negative sign to take into account that the frictional force will always be opposite to the direction of motion and hence to the direction of the velocity v. The value of β will depend on both the body — its size, its shape, and the roughness of its surface — and the density and viscosity of the fluid through which it is moving. The first step in solving eq. (4.15) is to relate a to v. Since $a = dv/dt$, we have

$$-\beta v = m\frac{dv}{dt}$$

We now have a differential equation for the velocity v as a function of the time t.[29] This differential equation can be solved for $v(t)$ by a method called separation of variables: we rearrange the equation so that everything involving the velocity variable is on one side and everything involving the time variable is on the other side, to obtain an equivalent equation that we can integrate:

$$-\frac{\beta}{m}\,dt = \frac{dv}{v}$$

If we denote the velocity at time $t = 0$ by v_0, then

$$-\int_0^t \frac{\beta}{m}\,dt = \int_{v_0}^{v} \frac{dv}{v}$$

$$-\frac{\beta}{m}\,t = \ln v\Big|_{v_0}^{v} = \ln v - \ln v_0 = \ln\frac{v}{v_0}$$

Raising both sides to the e to undo the log, we arrive at

$$\frac{v}{v_0} = e^{-\frac{\beta}{m}t}$$

$$v = v_0 e^{-\frac{\beta}{m}t}$$

This solution tells us that the initial velocity v_0 dies away exponentially, asymptotically approaching (but never at any finite time reaching) $v = 0$.

[29] A differential equation, as you might have guessed, is any equation that involves derivatives.

This is exactly what we would expect for a frictional force is proportional to the velocity: when the velocity is large, so is the frictional force, so that the body at first slows down rapidly. But as the body is slowed to a low velocity, the frictional force also becomes small, small enough that there is never enough friction to bring the body completely to rest.

Note also the dependence of our result for the velocity on the parameters m and β: for larger m, the $e^{-\beta t/m}$ and hence the velocity die away more slowly; for larger β, they die away more quickly. This makes perfect sense: the larger the mass m, the greater the body's inertia and therefore the harder to change its velocity, while the greater β the greater the frictional force.

For bodies moving through fluids, such as a baseball traveling through the air, the frictional force is nearly proportional to the square of velocity of the body. For bodies moving through air at speeds up to 100 mph, this is such a good approximation that it qualifies as almost nonbogus. If such a frictional force is the only force acting on the body, $F = ma$ becomes

$$\pm\beta v^2 = ma$$

where we have again written β for the (positive) constant of proportionality. This time, however, the signs are more complicated: since v^2 is always positive, there is no one sign that will make the frictional force opposite to the velocity for every case. If we assume that the body is moving in the positive direction, then we want friction to be in the negative direction and should therefore choose the negative sign in the above relation.

From this point, the solution is obtained by the same method as before: using $a = dv/dt$, separating variables, and integrating, we have

$$\begin{aligned} -\beta v^2 &= m\frac{dv}{dt} \\ -\frac{\beta}{m}\,dt &= \frac{dv}{v^2} \\ -\int_0^t \frac{\beta}{m}\,dt &= \int_{v_0}^{v} \frac{dv}{v^2} \\ -\frac{\beta}{m}\,t &= -\frac{1}{v}\bigg|_{v_0}^{v} = -\frac{1}{v} + \frac{1}{v_0} \end{aligned}$$

Solving for v, we arrive at

$$v = \frac{v_0}{1 + \frac{\beta}{m} v_0 t}$$

This solution tells us that the initial velocity v_0 again dies away quickly and again asymptotically approaches (but never at any finite time reaches) $v = 0$. Compared to the case of a frictional force proportional to the velocity, this v^2-proportional friction will cause the velocity to die away more quickly for

large v (since $v^2 \gg v$ for large v) and more slowly for small v (since $v^2 \ll v$ for small v). The dependence of the velocity on the parameters m and β is very similar to that of a frictional force proportional to the velocity, and for the same reasons.

So there you have it.

4.8 The Catenary

The catenary, shown in fig. (4.12), is the shape assumed by hanging cords of uniform linear density (mass per unit length), such as power lines between utility poles.[30]

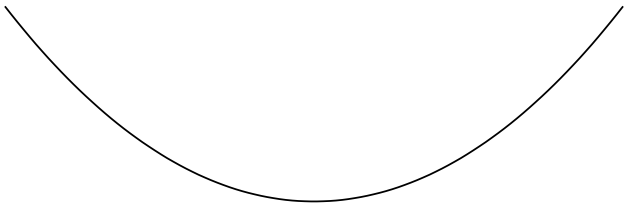


Figure 4.12: The Catenary

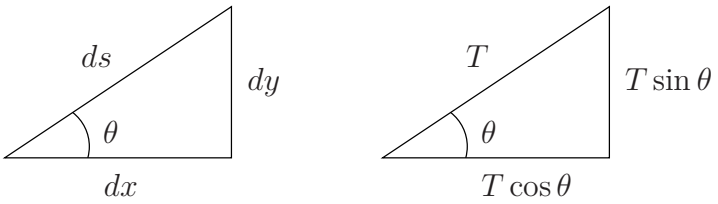


Figure 4.13: An Infinitesimal Segment of the Catenary

To see what happens physically, consider the infinitesimal segment ds of the catenary of linear density λ shown on the left side of fig. (4.13): the tension at the upper end of this segment will be greater than the tension at the lower end because the tension at the upper end has to support the added weight of the segment ds. More precisely, the *vertical component* of the tension at the upper end of this segment will be greater than the vertical component of the tension at the lower end by the weight $dm\, g = \lambda\, ds\, g = \lambda g\, ds$ of the segment; there being no additional horizontal forces for which to compensate, the horizontal component of the tension will remain constant. Setting the

[30] More precisely, it the shape assumed by cords that are not only of uniform density, but also infinitely flexible: to the extent that the cord is stiff, it will assume a shallower curve. But we won't let that bother us.

changes in the vertical and horizontal components of the tension equal to $\lambda g\, ds$ and zero, respectively, we have (see the right side of fig. (4.13))

$$d(T\sin\theta) = \lambda g\, ds \tag{4.16a}$$
$$d(T\cos\theta) = 0 \tag{4.16b}$$

What we would like to do is determine the geometric shape of the catenary analytically, in the form $y = y(x)$, starting from these two physical conditions, eqq. (4.16a) and (4.16b). From the geometry of fig. (4.13), you can see that

$$\tan\theta = \frac{dy}{dx} \tag{4.17a}$$
$$ds^2 = dx^2 + dy^2 \tag{4.17b}$$

We could try to proceed in the naive, direct way by using eqq. (4.17a) and (4.17b) to generate expressions for $\cos\theta$, $\sin\theta$, and ds that we could substitute into eqq. (4.16a) and (4.16b) — that would get us relations in terms of T, dx, and dy that we could in principle use to solve for $y = y(x)$ — , but the calculation we would be confronted with would turn out to be *very* ugly. Butt ugly, in fact. Wake-up-at-night-screaming ugly. Take our word for it. For starters, we would have to contend with the fact that the T that is left in our relations is not constant but is itself some as yet unknown function of x. Ugh.

This does, however, suggest a more elegant way to proceed: while the tension is not constant, its horizontal component $T_x = T\cos\theta$ is — that is what eq. (4.16b) is telling us. Let us therefore try writing eq. (4.16a) in terms of T_x rather than T:

$$T\sin\theta = T\cos\theta\,\tan\theta = T_x\tan\theta$$

Now we're in business: with this substitution, eq. (4.16a) becomes

$$d(T_x\tan\theta) = \lambda g\, ds$$
$$T_x\, d(\tan\theta) = \lambda g\, ds$$

where we have been able to pull T_x out of the differential because it is constant. And now we can sanely use eqq. (4.17a) and (4.17b) to rewrite this as

$$T_x\, d\left(\frac{dy}{dx}\right) = \lambda g\sqrt{dx^2 + dy^2}$$

which, if we divide both sides by dx, simplifies to

$$T_x\,\frac{d}{dx}\left(\frac{dy}{dx}\right) = \lambda g\,\frac{1}{dx}\sqrt{dx^2 + dy^2}$$

$$T_x \frac{d^2y}{dx^2} = \lambda g \sqrt{1 + \left(\frac{dy}{dx}\right)^2} \tag{4.18}$$

Eq. (4.18) is the differential equation that determines $y(x)$. Unfortunately, it's a nonlinear differential equation, and there are no general techniques for solving nonlinear differential equations as there are for solving linear differential equations. A nonlinear differential equation can be solved only in two cases: it falls into one of the classes of nonlinear differential equations to which people have managed to find the solutions, or you just happen to know ahead of time what the solution is. We will pursue the latter course here. It turns out that the general solution to eq. (4.18), as we will verify below by substitution, is a hyperbolic cosine function:

$$y = A \cosh \alpha(x + \xi) = A \frac{e^{\alpha(x+\xi)} + e^{-\alpha(x+\xi)}}{2}$$

where A, α, and ξ are parameters whose values we still need to determine. Since the effect of ξ is merely to shift the curve horizontally, without altering its shape at all, its value is of no concern to us, so for simplicity we will set $\xi = 0$. Our proposed solution then reduces to

$$y = A \cosh \alpha x = A \frac{e^{\alpha x} + e^{-\alpha x}}{2} \tag{4.19}$$

If we now note that [31]

$$\frac{d}{dx}(\cosh \alpha x) = \alpha \sinh \alpha x$$
$$\frac{d^2}{dx^2}(\cosh \alpha x) = \alpha^2 \cosh \alpha x$$
$$\cosh^2 \alpha x - \sinh^2 \alpha x = 1$$

then, substituting the proposed solution (4.19) into eq. (4.18), we have

$$T_x A \alpha^2 \cosh \alpha x = \lambda g \sqrt{1 + (A\alpha \sinh \alpha x)^2}$$

The only way that this relation can hold true for all x is if the square root on the right-hand side boils down to a $\cosh \alpha x$, and that will only happen if

$$A = \frac{1}{\alpha}$$

[31] Even if you aren't that familiar with hyperbolic functions, you should be able to verify these relations by taking derivatives of $\frac{1}{2}(e^{\alpha x} + e^{-\alpha x})$ and using the definition of the hyperbolic sine, $\sinh \alpha x = \frac{1}{2}(e^{\alpha x} - e^{-\alpha x})$.

So, using $A = 1/\alpha$, we have

$$\begin{aligned} T_x\left(\frac{1}{\alpha}\right)\alpha^2\cosh\alpha x &= \lambda g\sqrt{1+\left(\frac{1}{\alpha}\alpha\sinh\alpha x\right)^2} \\ T_x\alpha\cosh\alpha x &= \lambda g\sqrt{1+\sinh^2\alpha x} \\ &= \lambda g\cosh\alpha x \end{aligned}$$

which yields

$$\alpha = \frac{\lambda g}{T_x}$$

Eq. (4.19) therefore really is the solution for the catenary when we take

$$\alpha = \frac{1}{A} = \frac{\lambda g}{T_x}$$

With these values of A and α, our solution (4.19) becomes

$$y = \frac{T_x}{\lambda g}\cosh\left(\frac{\lambda g}{T_x}\,x\right) \tag{4.20}$$

Now, you might object, How is this a solution, when we don't know the value of T_x? One answer to this is, Who cares? All we care is that T_x is a constant, and, given that, we have determined the shape of the catenary in the analytic form $y = y(x)$.

The other answer is that we could, if we were so inclined, relate T_x to some other parameter we haven't yet considered, such as the total length of cord: between two utility poles, different lengths of wire can be strung. We expect intuitively that, if we were linemen stringing wire between poles, to take in slack and make the catenary more shallow we would have to pull on the wire, making it more taut and increasing the horizontal component of the tension. To see the exact relation between T_x and the length of cord, let's look at the simple case where the catenary is symmetric, that is, where the ends of the cord are at the same height (the case illustrated in fig. (4.12) on p.216): To get the length ℓ of cord that runs between utility poles at $x = -a$ and $x = a$, we simply sum up the infinitesimal arc lengths ds as we go between poles:

$$\begin{aligned} \ell &= \int_{-a}^{a} ds \\ &= \int_{-a}^{a}\sqrt{dx^2+dy^2} \\ &= \int_{-a}^{a} dx\,\sqrt{1+\left(\frac{dy}{dx}\right)^2} \end{aligned}$$

$$
\begin{aligned}
&= \int_{-a}^{a} dx \sqrt{1 + \left(\frac{d}{dx} \left(\frac{T_x}{\lambda g} \cosh\left(\frac{\lambda g}{T_x} x \right) \right) \right)^2} \\
&= \int_{-a}^{a} dx \sqrt{1 + \sinh^2\left(\frac{\lambda g}{T_x} x \right)} \\
&= \int_{-a}^{a} dx \, \cosh\left(\frac{\lambda g}{T_x} x \right) \\
&= \frac{T_x}{\lambda g} \sinh\left(\frac{\lambda g}{T_x} x \right) \Bigg|_{-a}^{a} \\
&= 2 \frac{T_x}{\lambda g} \sinh\left(\frac{\lambda g}{T_x} a \right)
\end{aligned} \tag{4.21}
$$

While this is a transcendental equation that cannot be solved in closed form, in principle it gives T_x in terms of ℓ, a, λ, and g.

And that's about all we're going to say about catenaries.

But now, every time you look at power or phone lines, you'll think *catenary* and *hyperbolic cosine*. Either that, or you'll start shrieking and sobbing uncontrollably and really freak out the people you're with.

4.9 Problems

1. While you are patiently waiting in an interminably long line in the dining hall, you are rudely pushed from behind by an oaf.

   (a) If your mass is 60 kg and the oaf pushes you with a 360 N force, what is your acceleration?

   (b) If you are a pacifist and do not respond, do you exert any force on the oaf?

   (c) If instead your weight is 140 lb and the oaf pushes you with a 280 lb force, what is your acceleration? (If you are crafty, there is no need to convert pounds to Newtons.)

   (d) In reality, you do not usually experience motion as a result of being pushed in line. Why is this?

2. A bullet from a .357 magnum (mass 158 grains = 10.2 g) strikes a Barney doll (65 kg) at 440 m/s and comes to rest as it penetrates 12 cm. What average force do the bullet and doll exert on each other as it is brought to rest?

3. A bullet from a .357 magnum (mass m_b) strikes a Barney doll (mass m_y) at speed v_b and comes to rest as it penetrates to a depth ℓ. What average force do the bullet and doll exert on each other as it is brought to rest?

4. (a) A dead cat has a mass of 5.0 kg. How much would this cat weigh on the Moon, where gravity is only $\frac{1}{6}$th what it is on Earth?

   (b) If a dead cat weighs 4.0 lb on the Moon, what is its mass?

   (c) Actually, the mass and radius of the Moon are 7.347673×10^{22} kg and 1.7360×10^6 m, and those of the Earth are 5.9723×10^{24} kg and 6.378140×10^6 m. Determine a more precise result for the ratio g_{moon}/g_{earth}.

5. During an amicable disagreement, your best friend exerts a horizontal 150 lb force as he or she gives you a noogie while smashing your head against a wall.

   (a) What are the magnitudes and directions of all the horizontal forces acting on your head?

   (b) What are the magnitudes and directions of all the horizontal forces acting on the wall?

   (c) What are the magnitudes and directions of all the horizontal forces acting on your friend's fist?

   (d) Which pairs of these forces are related to each other by Newton's third law?

   (e) Which pairs of forces are related to each other by Newton's second law?

6. You hold a container of General Tso's (mass m_{tso}) at rest in the palm of your outstretched hand (mass m_{hand}).

   (a) What are all the forces acting on the container?

   (b) What are all the forces acting on your hand?

   (c) Determine the values of all of these forces.

   (d) Which pairs of forces are related to each other by Newton's third law?

   (e) Which pairs of forces are related to each other by Newton's second law?

7. After failing out of college, you get a job with as the circus clown who gets shot out of a cannon. While you are airborne, do you feel weightless? Explain.

8. As a crowded elevator stops at a floor, a huge feckless oaf of mass m_o squeezes in and stands on your foot.

   (a) What does feckless mean? And why are oafs almost always feckless? Or should that be *oaves*?

   (b) Assuming that the elevator and everyone in it is remaining at rest, draw force diagrams for yourself (who we will take to be mass m_y) and for the oaf. (Although the oaf is standing on your foot, physically the situation is the same as if the oaf were standing on your head: if you are represented by a dot in the force diagram, it does not matter on which part of you the oaf is standing; either way, you are supporting the oaf.)

   (c) How will the force diagrams change if the elevator has a nonzero velocity or a nonzero acceleration?

   (d) Solve for all of the forces in the case that the elevator is

      i. Sitting at rest.
      ii. Moving upward at a constant *speed* v_u
      iii. Moving downward at a constant *speed* v_d.
      iv. Accelerating upward at $\frac{1}{3}g$.
      v. Accelerating downward at $\frac{1}{3}g$.
      vi. Accelerating downward at g.

   (e) Descriptively, what might the elevator be doing in each of the cases in #8d? That is, could it be moving up or down, and could it be starting up or coming to a stop? How does this gibe with what you feel when you ride an elevator?

   (f) How would it be possible to give the elevator a downward acceleration of $2g$? If the elevator did have a downward acceleration of $2g$, what would happen to you and the oaf?

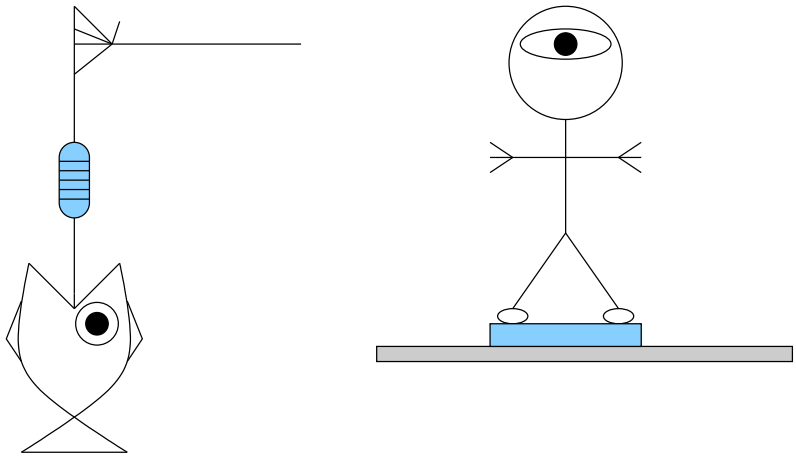


Figure 4.14: Problem 9

9. (a) The left side of fig. (4.14) shows a fish being weighed on a spring scale: the fish (mass m) is hung from the bottom of the scale, which in turn is held up by a person with an unusual five-fingered hand with an opposable thumb.

   Determine all of the forces acting on the scale, assuming that the scale itself is massless, and compare these forces to the reading you expect the scale to give.

   (b) The right side of fig. (4.14) shows a more normal three-fingered person (mass m) standing on the most hated piece of technology in America: the bathroom scale.

   Determine all of the forces acting on the scale, assuming that the scale itself is massless, and compare these forces to the reading you expect the scale to give.

   (c) What would happen to the values of these forces, and what would the scales therefore read, if they were being used in an elevator with a downward acceleration of $\frac{1}{3}g$?

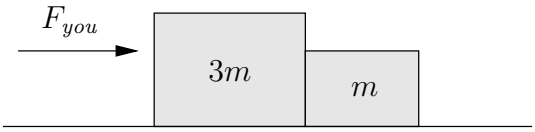


Figure 4.15: Problem 10

10. To earn some community-service credit over the summer, you get an internship with the mafia and help illegally dump medical waste off the Jersey shore. At one point, you are pushing two crates of waste, of masses $3m$ and m, down a dock by applying a horizontal force F_{you}, as shown in fig. (4.15). The dock has been greased and is therefore virtually frictionless.[32]

    (a) Determine

        i. The acceleration of the crates.

        ii. The force between crates.

        iii. The net force on each crate.

    (b) You should see a very sensible relationship among some of your results for the preceding parts of the problem. If not, keep looking.

    (c) How can you push crates down a frictionless dock?

    (d) What if the crates were reversed, so that you were pushing on the crate of mass m? Which of your results would change and which remain the same?

[32] Okay, so at this point the problem becomes a little unrealistic.

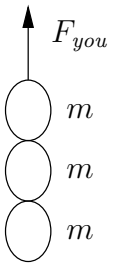


Figure 4.16: Problem 11

11. You pull with an upward force F_{you} on the topmost link of a rather lame chain consisting of just three links, each of mass m, as shown in fig. (4.16).

    (a) Determine (not necessarily in this order)

        i. The net force on each link.
        ii. The acceleration of the chain.
        iii. The force exerted on each link by the link (if any) below it.
        iv. The force exerted on each link by the link (if any) above it.

    (b) Suppose the chain instead consisted of 1,000,000 links. Determine (not necessarily in this order)

        i. The net force on the 987,654th link from the top.
        ii. The acceleration of the chain.
        iii. The force exerted on the 987,654th link by the link below it.
        iv. The force exerted on the 987,654th link by the link above it.

12. Draw a force diagram for each of the following utterly lame situations:

    (a) A mass m sliding down a frictionless plane inclined at angle θ to the horizontal.

    (b) A mass m sliding up a frictionless plane inclined at angle θ to the horizontal.

    (c) A mass m sliding down a plane inclined at angle θ to the horizontal when there is friction.

    (d) A mass m sliding up a plane inclined at angle θ to the horizontal when there is friction.

    (e) Two masses, m_1 on top of m_2, sliding down a plane inclined at angle θ to the horizontal, with friction both between m_2 and the plane and between m_1 and m_2. Mass m_1 is fixed on top of m_2 (that is, m_1 is not slipping on m_2).

13. (a) You are kidnapped by vicious Bokononist space aliens who take you to a barren distant planet, force you to put on a bunny suit, and leave you at rest in the center of a perfectly frictionless and level frozen pond of ice-nine. Is there any way that you can get off the pond and escape?

    (b) You are floating around way out in the void of space. Fortunately, you are in a space suit. Unfortunately, floating next to you is a box full of Spice-Girl CDs. You and the box are at rest relative to each other. If you push the box away from you, what happens?

    If you get this but didn't get #13a, this would be a good time to go back and look at #13a again.

    (c) Suppose you are trying to push a trunk containing a corpse. If, by Newton's third law, the trunk always exerts a backward force on you equal in magnitude to the push that you are applying, then these two third-law forces should always cancel each other out and you should never be able to move the trunk.

    Explain why this reasoning is boneheaded. Be precise.

14. If you drop two objects, one solid, the other hollow, but both made of the same substance and otherwise identical, which will hit the ground first? Prove your assertion.

15. After waiting forever in the lunch line, you discover that it was Buffalo-wing day and that there is only a single Buffalo wing left in the bottom of the tray. You (65 kg) and another student (100 kg), each armed with tongs, grab hold of the 30 g wing. The other student pulls on the wing with a force of 120 N in the positive x direction; you pull on it with a 200 N force at a counterclockwise angle of 135° with the positive x axis.

    (a) Are forces of 120 N and 200 N of realistic magnitude? What are their values in pounds (lbs)? Or if we think of hefting objects, what masses (in kg) would have equivalent weights?

    (b) If we neglect the wing's weight, what are the magnitude and direction of the net force on the wing?

    (c) Was it reasonable to neglect the wing's weight in the preceding part?

    (d) What are the magnitude and direction of the wing's acceleration?

    (e) What speed would the wing acquire if it experienced this acceleration for a mere $\frac{1}{10}$ sec?

    (f) In reality, Buffalo wings don't acquire these kinds of velocities (although lunch would arguably be much more entertaining if they did). Why is this?

16. In this problem, we will use the bogus "law" of friction $F_{fric} = \mu N$. Be sure to draw a force diagram for each part to help you see what is going on.

    One day your butt falls off, so, following your coach's advice, you put it in a paper bag and drag it to practice with you.[33]

    (a) If your butt weighs 250 N and it takes a horizontal pull of 100 N to get your butt moving across level ground, what is the coefficient of static friction between your butt and the ground?

    (b) If the coefficient of kinetic friction between your butt and the ground is 0.20, what horizontal pull must be applied to keep your butt moving along at a constant velocity?

    (c) If instead a 100 N horizontal pull continues to be applied to your butt, what is its acceleration?

    (d) If you (102 kg, buttless) sit on your butt while a friend pulls it along, what horizontal pull must your friend now apply to your butt to get it moving?

    (e) What if instead your friend pushes horizontally from behind?

    (f) What if your friend pushes from behind, but downward at an angle of 30° to the horizontal?

    (g) What if your friend pulls from the front, but upward at an angle of 30° to the horizontal?

    (h) Why (physically) is your answer for #16g smaller than that for #16f?

[33] Coaches are always thundering about the life-or-death need to make practice. *"I don't care if your butt falls off — put it in a paper bag and bring it with you."* All coaches have said this at one time or another. You have to remember that coaching is a form of mental illness.

17. (Based on the experience of an acquaintance of the author.) Under cover of darkness, you (mass m) drag a corpse (mass m') up an incline so that you can toss it into a ravine. The incline is at angle θ to the horizontal with coefficients μ_s and μ_k of bogus static and kinetic friction, respectively.

    (a) If you pull *parallel to the incline*, how much force must you apply

        i. To get the body moving?

        ii. To keep the body moving at a constant speed? Think before you start calculating.

    (b) If instead you push the body from behind with a *horizontal* force, how much force must you apply

        i. To get the body moving?

        ii. To keep the body moving at a constant speed? Think before you start calculating.

    (c) Make physical sense of the dependence of your results on the parameters θ and μ_s (or μ_k): how do your results behave when μ is large or small or when θ is close to zero or $\frac{\pi}{2}$? Be sure to analyze any other special values of these parameters (that is, values at which something physically significant occurs) as well.

18. Some stupid block slides down a frictionless incline at angle θ to the horizontal. But, get this: the incline is inside an elevator! Cool, huh?

    (a) Okay, so maybe it's not all that cool. But you still have to determine the acceleration of the block *relative to the incline* when the elevator is

        i. At rest.

        ii. Moving upward at a constant speed v.

        iii. Moving downward at a constant speed v.

        iv. Accelerating upward with an acceleration of magnitude a.

        v. Accelerating downward with an acceleration of magnitude a.

        vi. Accelerating downward at g.

        You may want to think in terms of the *effective* value of g in each of these cases.

    (b) Unless you are a pain freak, you worked out #18a in terms of the effective value of g. Now prove that this method is valid by deriving it from $\mathbf{F} = m\mathbf{a}$ for the block. You may want to set things up in terms of vectors and to split the acceleration of the block into two pieces: the acceleration $\mathbf{a}_e$ that it shares with the elevator and the acceleration $\mathbf{a}_i$ that it experiences relative to the incline.

19. To test the effectiveness of prototype baby powders, the butts of test babies are powdered and then placed on inclines made of rough, splintery wood. It is found that the most smoothly powdered baby butt slides down an incline of length ℓ at angle θ to the horizontal with a constant acceleration of $\frac{1}{3}g\sin\theta$.

    (a) If a lab technician, too lazy to carry the baby back to the starting point, simply slides the baby back up the incline, what initial velocity must the technician give the baby in order that it make it all the way back up?

    (b) Exactly what do we know about the nature of the frictional force from the given information?

    (c) Realistically, how would you know that the acceleration was $\frac{1}{3}g\sin\theta$ to begin with?

20. You want to dump a truckload of sand (or gravel or mulch or anthrax or whatever) of volume V, and you want the shape of the heap in which you dump it to minimize the space (area) taken up on the ground. The coefficient of friction between particles in the load is μ. Give a compelling argument that the heap should be conical, with the ratio of height to the radius of the base being equal to μ.

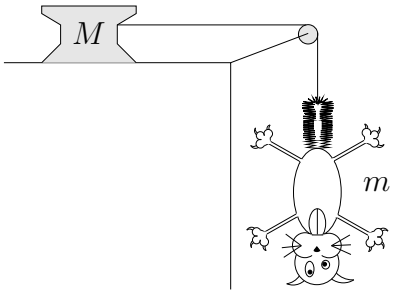


Figure 4.17: Problem 21

21. Fig. (4.17) is supposed to show a cat connected to an anvil by a massless cord slung over a pulley. The table in fig. (4.17) is frictionless and level and the pulley is massless and frictionless.

    (a) Determine the acceleration of the cat and the tension in the cord.

    (b) Why in real life do you not get any motion in an arrangement like this?

    (c) Explain how your results for the tension and acceleration make sense in the limits

        i. $M \gg m$.

        ii. $m \gg M$.

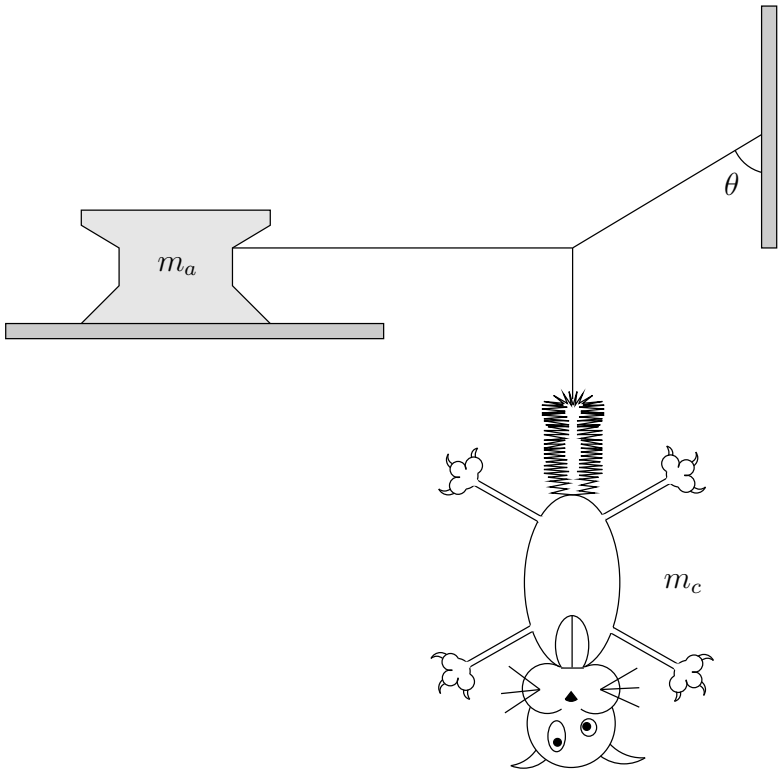


Figure 4.18: Problem 22

22. Fig. (4.18) shows a cat suspended in an admittedly very artificial manner that could not possibly serve any useful purpose.

    (a) Notwithstanding, determine the minimal coefficient of bogus friction that will prevent the anvil from slipping. (Note that at the junction of the massless cords there is a knot that is, of course, also massless. There are thus effectively three separate cords, each with its own tension.)

    (b) Make physical sense of your result for the various limits of the parameters m_c, m_a, and θ.

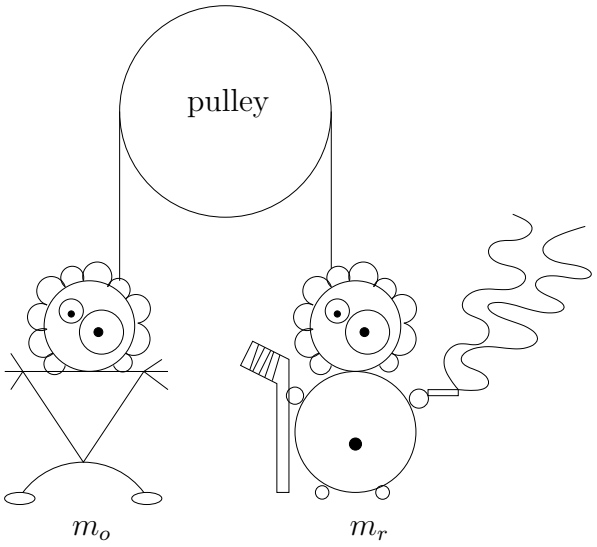


Figure 4.19: Problem 23

23. Fig. (4.19) shows a classic Atwood's machine. Or rather, a classic Atwood's machine with the more usual totally lame block masses replaced by two colleagues of the author as infants. With our usual callous disregard of reality, we will take the rope to be massless and the pulley to be both massless and frictionless.

    (a) Draw force diagrams for each of the two masses and for the pulley. (Remember that the pulley will be mounted on some sort of axle and that this axle can exert a force on the pulley.)

    (b) Set up $F = ma$ for each of the two masses. Be mindful that when m_o goes down, m_r goes up and vice versa.

    (c) Solve for the acceleration of each mass and for the tension in the rope.

    (d) Explain how your results for the tension and acceleration make sense in the limits

        i. $m_r \gg m_o$ (or vice versa).

        ii. $m_r = m_o$.

    (e) Explain how your result for the acceleration makes sense when the two masses are regarded *as a system*.

    (f) What force is exerted on the pulley by its axle?

    (g) Who the $****$ was Atwood, and what possible use is this contraption? (We may never figure out who Atwood was, but there are in fact common uses for this thing — those of you who live in old houses will almost certainly have one sort of Atwood's mechanism at home, and those of you who live in old mansions may have a second, more entertaining sort.)[34]

[34] Actually, Atwood's machine is of some historical interest in the development of the laws of mechanics: for everything you ever wanted to know about Atwood's machine and then some, see S. Esposito and E. Schettino, *Spreading scientific philosophies with instruments: the case of Atwood's machine*, `https://arxiv.org/abs/1204.2984`.

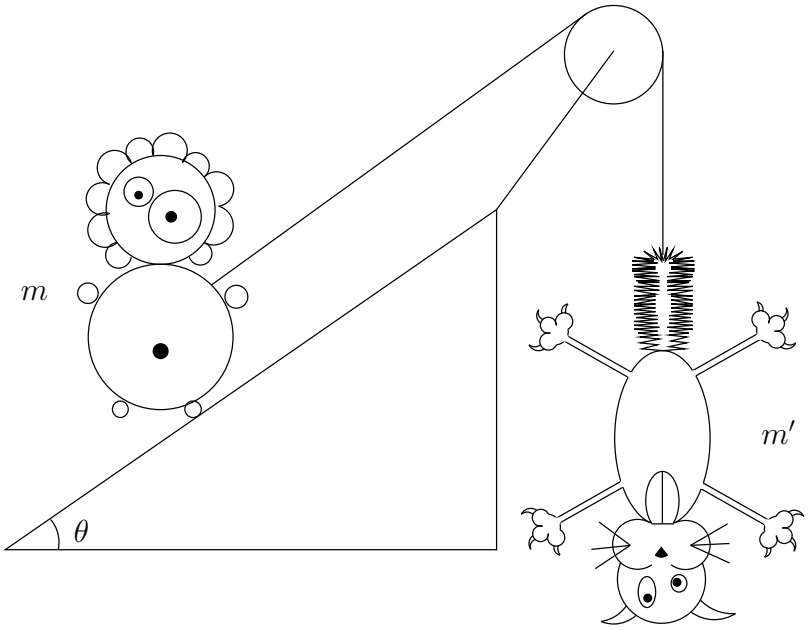


Figure 4.20: Problem 24

24. (a) If the incline in fig. (4.20) is frictionless,
    i. Determine the tension in the cord and the acceleration of the masses.
    ii. Show that there are limits of the values of the parameters m, m', and θ that will reproduce your results for #21 and #23, and make physical sense of these limits.

    (b) If instead there is bogus friction between the baby and the incline,
    i. Determine the minimal coefficient of static friction that will prevent motion.
    ii. Make physical sense of your result in the limits $\theta \to 0$ and $\theta \to \frac{\pi}{2}$.

    (c) If the coefficient of bogus kinetic friction between the baby and the incline is μ_k, determine the tension in the cord and the acceleration of the masses if
    i. The cat is descending.
    ii. The cat is ascending. Think before you calculate.

    (d) Your results for the acceleration in #24a and #24c should seem almost self-evident if you look at the two masses as a system. If not, keep looking.

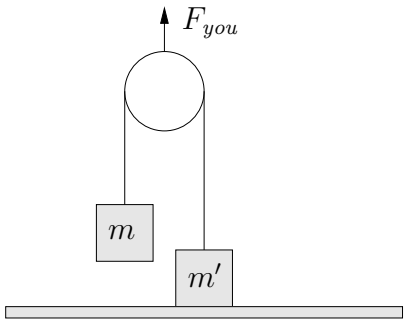


Figure 4.21: Problem 25

25. In fig. (4.21) the mass m' is in contact with the surface and the round thingie is the usual massless, frictionless pulley. The pulley is not mounted; it is in midair and held up solely by an upward pull F_{you} that you are (through the axle or whatever) exerting on it.

    (a) Solve for the accelerations of the mass m and of the pulley if you are *not* pulling hard enough to lift m' off of the surface. See the footnote if you need a hint.[35]

    (b) Determine the minimal force F_{you} that will cause m' to lose contact with the surface.

    (c) i. Treating F_{you} as a given, determine the accelerations of m, m', and (if you can) the pulley when m' has lost contact with the surface. See the footnote if you need a hint.[36]

    ii. This is getting a bit ahead of the game, but verify that if you regard the pulley and masses as a system, then

    $$\mathbf{F}_{net} = (m + m')\mathbf{a}_{cm}$$

    where $\mathbf{F}_{net} = \mathbf{F}_{you} - (m + m')\mathbf{g}$ is the net force on the system and, as we will see in Chapter 6, the acceleration $\mathbf{a}_{cm}$ of the center of mass is given, in terms of the accelerations $\mathbf{a}$ and $\mathbf{a}'$ of m and m', by

    $$\mathbf{a}_{cm} = \frac{1}{m + m'}(m\mathbf{a} + m'\mathbf{a}')$$

    (d) What special cases can you use to check whether your results make sense?

    (e) What features of this problem make its analysis different from that of Atwood's machine?

[35] Don't forget to set up $F = ma$ for the pulley, and when you do so note carefully how the cord is connected to the pulley and how the tension in the cord therefore acts on it. Looking at how the cord winds over the pulley will also allow you to solve for the pulley's acceleration.

[36] To get the pulley's acceleration, think about how far the pulley rises in terms of the distances that m and m' rise, taking into account how the cord winds around the pulley.

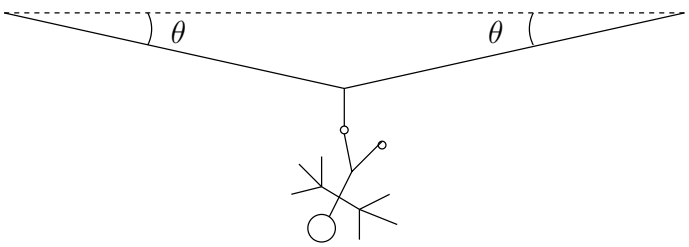


Figure 4.22: Problem 26

26. Suppose that while out mountaineering you (mass m) find yourself in the unfortunate situation shown in fig. (4.22). At the Y-shaped junction there is a knot, which means that there are effectively three separate ropes, each with its own tension.

    (a) What are the tensions in these three segments of rope? Note that when tied together, massless ropes will yield massless knots.

    (b) Make physical sense of the dependence of these tensions on θ.

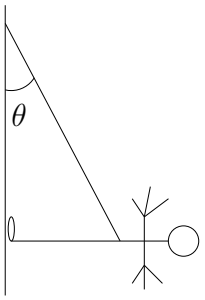


Figure 4.23: Problem 27

27. The ludicrously poor fig. (4.23) shows a mountain-climbing technique known as *rappelling*. Suppose you find yourself (mass m) in this ridiculous situation, with your legs and body at a right angle to the rock face.

    (a) Draw a force diagram for yourself, assuming that there is friction between your feet and the rock face.

        i. What can you conclude about the horizontal and vertical forces the rock face exerts on you?

        ii. Is it possible to solve for the forces acting on you? If not, why not?

    (b) Suppose the rock face is instead frictionless.

        i. What can you conclude about the horizontal and vertical forces the rock face exerts on you?

        ii. Is it possible to solve for the forces acting on you? If not, why not?

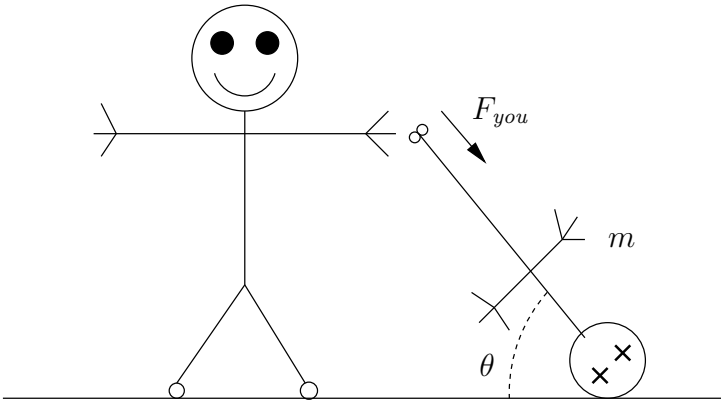


Figure 4.24: Problem 28

28. After your roommate is electrocuted trying to clean up spilled soda with a vacuum cleaner, you discover that his or her body, now stiff as a board, actually makes an excellent mop. You therefore proceed to mop up the floor with your roommate (mass m), applying a push F_{you} along (that is, parallel to) his/her body, which is at angle θ to the floor. There is bogus friction with static and kinetic coefficients μ_s and μ_k, respectively, between your roommate's head and the floor.

    (a) Determine the minimal force F_{you} needed to get your roommate moving.

    (b) Determine the force F_{you} needed to keep your roommate moving at a constant speed across the floor.

    (c) What values or limits of the parameters m, g, μ_s, μ_k, and θ are of physical interest in your two results for the force F_{you}? Make physical sense of your results for these values and limits.

    (d) Certainly it would make sense to consider the cases $\theta = 0$, $\theta = \frac{\pi}{2}$, θ small, or θ close to $\frac{\pi}{2}$. We're not saying that anything interesting does or doesn't happen in these cases, merely that one could sensibly examine them. Explain why each of the following cases would or would not be sensible to consider:

        i. Small or large m.
        ii. $m = 1$.
        iii. Small or large μ_s.
        iv. $\mu_s = 1$.
        v. Small or large g.
        vi. $g = 1$.

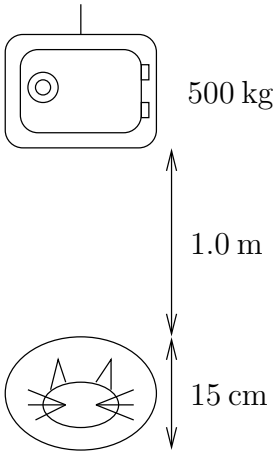


Figure 4.25: Problem 29

29. In what is very possibly the best illustration we've ever drawn (or are likely ever to draw), fig. (4.25) shows a 500 kg safe dangling 1.0 m above a sleeping cat by a spindly little thread, which of course snaps. Determine the force exerted on the cat by the safe as it smashes the cat, originally 15 cm high, into a disgusting 1.0 cm-thick film, and do this preferably in terms of the variables $m = 500\,\text{kg}$, $\ell = 1.0\,\text{m}$, $h_0 = 15\,\text{cm}$, and $h = 1.0\,\text{cm}$. Assume that the acceleration of the safe is constant as it squashes the cat.

30. (a) Way back in your parents' day, there was a device known as a phonograph that played music on 12 inch-diameter (30.5 cm) LP records rotating at $33\frac{1}{3}$ rpm (revolutions per minute).

    i. At what speed is an ant standing on the edge of an LP moving when it is rotating at $33\frac{1}{3}$ rpm?

    ii. How many g's of centripetal acceleration does the ant then experience?

    iii. Would the ant have trouble holding on?

    iv. If it takes the LP 0.25 sec to "spin up" from rest to $33\frac{1}{3}$ rpm, how many g's of average tangential acceleration does the ant experience during this acceleration?

    (b) By contrast, the 3 inch-diameter (7.6 cm) disk of your hard drive likely rotates at 7200 rpm.

    i. At what speed would an ant standing on the edge of the disk be moving?

    ii. How many g's of centripetal acceleration would the ant experience?

    iii. Would the ant have trouble holding on?

31. On a roller coaster ride you (mass m) go around a circular loop-the-loop of radius R at speed v.

    (a) Determine the number of g's you experience at the top and bottom of the loop. Remember that your perceived weight is determined by the normal force.

    (b) Physically interpret the behavior of your solution as a function of v, R, and m.

    (c) What is happening physically in the special cases

        i. $v > \sqrt{gr}$?

        ii. $v < \sqrt{gr}$?

        iii. $v = \sqrt{gr}$?

    (d) In reality, your speed varies as you go around the loop: you slow down as you fight gravity on the way up and speed up as gravity pulls you on the way down. We won't be able to deal with this complication until we get to the chapter on energy, but it turns out that in order for you to make it around the top of a circular loop-the-loop, your speed would have to be such that the number of g's you experience at the bottom would be causing people to lose consciousness, have strokes, etc. While this would make for very entertaining photos, unfortunately there are other considerations, and so real loop-the-loops are not circles but clothoid loops that are sort of egg-shaped, with the small end of the egg pointing up. The effect is that the radius of curvature is smaller at the top of the loop, to compensate for your slower speed.

        But for the simple bumps and dips you ride over in a roller coaster there are no such complications to worry about. So: What realistic values for the normal force between you and the seat would give a good ride over these bumps and dips? Consider just two points: when you are rounding the bottom of a dip or going over the top of a hump (both of which you may treat as circular arcs).

32. You weigh yourself on a scale when you are at the Equator and then weigh yourself again on the same scale when you are at the North Pole. What, according to the scale, will the ratio of your weight at the Equator to your weight at the North Pole be? Work out a result in terms of variables, then work out the numbers. Do not worry about the Earth's equatorial bulge; assume that it is a sphere of radius 6.378140×10^6 m.

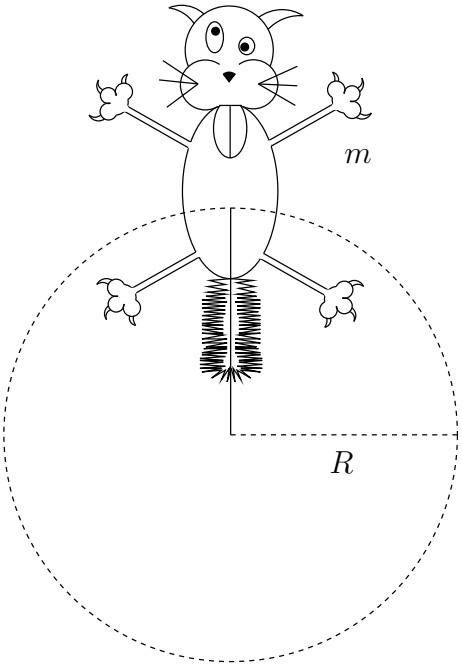


Figure 4.26: Problem 34

33. On an uneventful day you pass the time by twirling a blasting cap on the end of a string in a vertical circle. When the blasting cap is at the very top of the circle, the string suddenly snaps. Sketch and describe in words the subsequent motion of the blasting cap.

34. A standard test cat of mass m is lassoed and swung in a *vertical* circle of radius R, as shown in fig. (4.26).

    (a) If the cat is moving at speed v at the bottom of the circle, what is the tension in the rope at that point?

    (b) If instead the rope just barely goes slack at the top of the circle, how fast is the cat moving at that point?

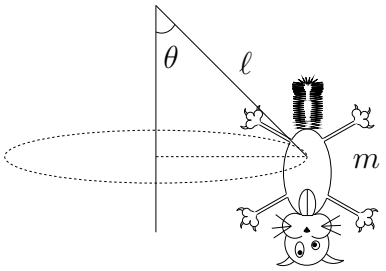


Figure 4.27: Problem 35

35. The test cat of #34 is now used to play tetherball.[37] At the moment shown in fig. (4.27), the cat is swinging around the pole in a *horizontal* circle at the end of a rope of length ℓ at angle θ to the vertical.

    (a) What is the tension in the rope and how fast is the cat moving?

    (b) Make physical sense of your results in the special cases

        i. $\theta \to 0$.

        ii. $\theta \to \frac{\pi}{2}$.

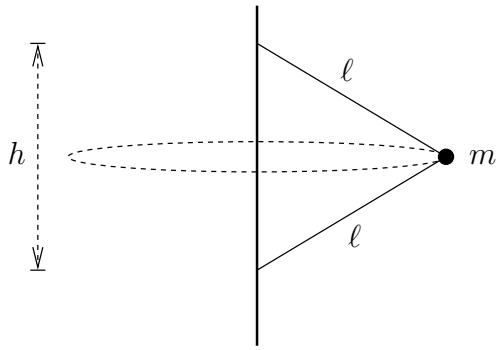


Figure 4.28: Problem 36

36. In fig. (4.28), a mechanical arrangement known as a governor, the mass m is rotating in a horizontal circle around the vertical axis.[38]

    (a) If the angular frequency of the mass's revolution is ω, determine the tensions in the upper and lower cords.

    (b) What limits are there on the possible values of m, ℓ, h, or ω?

[37] It is strongly recommended that novices first master this game with a declawed cat.

[38] To see just what purpose a governor serves, we will have to wait until we have done rotational motion and dealt with moments of inertia and angular momentum. But believe it or not it does do something useful.

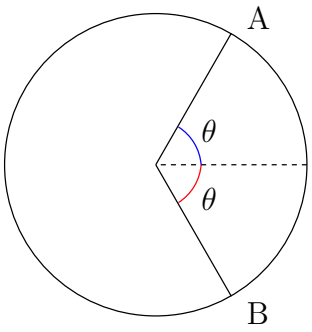


Figure 4.29: Problem 37

37. Fig. (4.29) shows a sharp pointy rock of mass m, tied to a string, being swung in a *vertical* circle of radius R for sinister purposes.

    (a) Determine the tension in the string and the rock's tangential acceleration if the speed of the rock is v_A at point A (which is θ above the horizontal).

    (b) Determine the tension in the string and the rock's tangential acceleration if the speed of the rock is v_B at point B (which is θ below the horizontal).

    (c) Make physical sense of your results when

        i. $\theta \to 0$.

        ii. $\theta \to \frac{\pi}{2}$.

        iii. $\theta \to -\theta$.

38. The turns of the track in the Field House are banked.[39] Estimate the running speed for which the turns are designed and calculate the ideal banking angle for that speed. Does your result for the banking angle gibe with what you see in the Field House? (You should also be able to come up with a reasonable estimate of the banking angle of the track in the Field House by thinking about the geometry of the triangle of which the track surface is the hypotenuse.)

39. In our discussion of road banking in §4.4.1, we considered only the case of ideal banking, when no friction is required to make the turn.

    (a) If you take the turn too quickly (that is, at a higher speed than that for which it is banked), you would, in the absence of friction, tend to slide up the banking. And if you take the turn too slowly, you would, in the absence of friction, tend to slide down the banking. Explain physically why this is so. (You will want to think in terms of the force relations.)

[39] At least, in our field house they are banked, and we are assuming that is commonly the case with indoor tracks.

(b) For the case that you are taking the turn too quickly:

i. Modify the force diagram in fig. (4.8) on p.207 to include friction.

ii. Show that the force relations, with a little algebra, yield

$$F_f = \frac{mv^2}{r}\cos\theta - mg\sin\theta$$

for the frictional force when you take the turn too quickly but without slipping.

iii. Show that the force relations, with a little algebra, yield

$$\frac{v^2}{gr} = \frac{\mu\cos\theta + \sin\theta}{\cos\theta - \mu\sin\theta}$$

for the maximal speed at which you can take the turn without slipping.

(c) Repeat #39b for the case that you are taking the turn too slowly and get a relation for the minimal speed at which you can take the turn without slipping.

(d) Explain how your result for #39(b)ii does or does not make sense in the following cases:

i. Large or small v.

ii. Large or small r.

iii. $\theta \to \frac{\pi}{2}$.

iv. $\theta \to 0$.

(e) Explain how your result for #39(b)iii does or does not make sense in the following cases:

i. $\cos\theta > \mu\sin\theta$.

ii. $\cos\theta < \mu\sin\theta$.

iii. $\cos\theta - \mu\sin\theta \to 0$.

iv. $\theta \to 0$.

v. $\theta \to \frac{\pi}{2}$.

(f) Explain how your result for #39c does or does not make sense in the following cases:

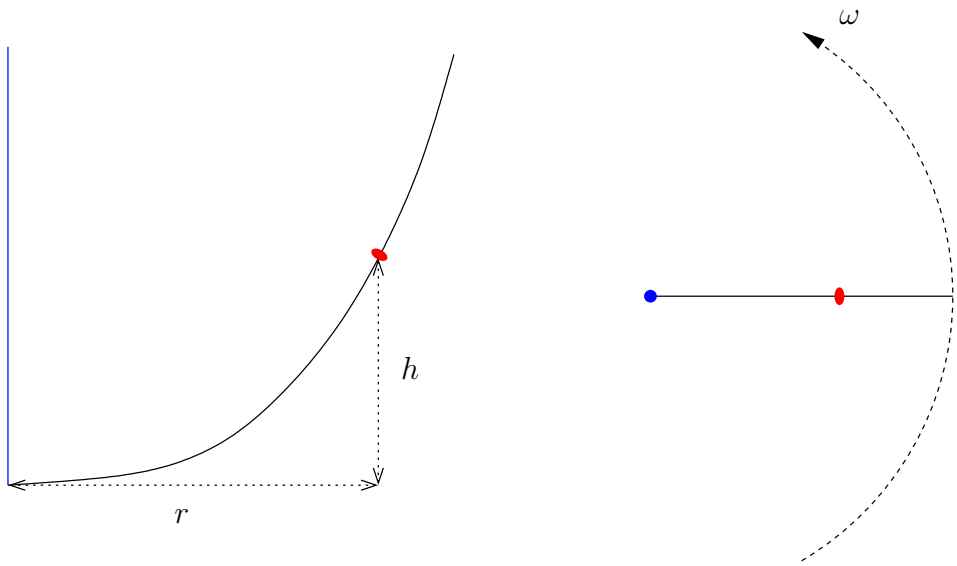
i. $\sin\theta > \mu\cos\theta$.

ii. $\sin\theta < \mu\cos\theta$.

iii. $\sin\theta - \mu\cos\theta \to 0$.

iv. $\theta \to 0$.

v. $\theta \to \frac{\pi}{2}$.



View from the side View from above

Figure 4.30: Problem 40

40. Fig. (4.30) shows a piece of wire (the solid black curve) bent upward into a curve $h(r)$, where r is the radial distance, and h the height, from the bottom end of the wire. An essentially point-like bead of mass m (shown in red) is mounted on the wire and free to slide frictionlessly along it. The wire is then rotated around a vertical axis (shown in blue) at a constant angular velocity ω.

    (a) Determine the height h at which the bead will be in equilibrium, with the forces on it providing for its circular motion, if the curve of the wire is the straight line $h = cr$, where c is a positive constant.

    (b) Determine the height h at which the bead will be in equilibrium, with the forces on it providing for its circular motion, if the curve of the wire is $h = cr^3$, where c is a positive constant (though of course different from the c of the preceding part).

    (c) Determine the height h at which the bead will be in equilibrium, with the forces on it providing for its circular motion, if the curve of the wire is $h = cr^2$, where c is a positive constant. You may find your result surprising at first, but on reflection it should make sense. If not, keep reflecting.

41. (a) Use Newton's law of gravity to estimate the gravitational attraction that you and the person sitting next to you exert on each other.

    (b) If gravity is so weak, why do you seem to be held down to the ground by such a substantial force?

42. If it is in a circular orbit 360 km above the Earth's surface, how long does it take that incredible waste of research funds known as the International Space Station to complete each full revolution around the Earth? The mass and radius of the Earth are 5.9723×10^{24} kg and 6.378140×10^{6} m.

43. How high above the surface of the Earth must a satellite's circular orbit be if it is to be in a geosynchronous orbit over the equator, that is, in an orbit with a period of 24 hours (which will keep the satellite perpetually over the same point on the Earth)? The mass and radius of the Earth are 5.9723×10^{24} kg and 6.378140×10^{6} m.

44. Show that it is not possible to orbit the Earth at 8000 m/s.

45. Prove Kepler's third law for circular orbits. (Note that for a circle, the semimajor and semiminor axes are both simply the radius.)

46. The Moon, as you may be aware, orbits around the Earth, and Phobos orbits around Mars. If the radius of the Moon's orbit is 40 times that of Phobos and it takes the Moon 80 times as long as Phobos to complete a full orbit, how many times more massive is the Earth than Mars? (Treat both orbits as circular — which, as it turns out, is actually not a bad approximation.)

47. Justify your answers to each of the following by calculation or reasoning:

    (a) In gravitational orbits,

        i. The mass of the satellite always matters.
        ii. The mass of the satellite never matters.
        iii. The mass of the satellite sometimes matters, sometimes doesn't.
        iv. Déjà vu can sometimes occur.
        v. Déjà vu can sometimes occur.

    (b) If a satellite is changed to an orbit in which it travels twice as fast, the period of its orbit will change by a factor of

        i. Four.
        ii. Two.
        iii. One half.
        iv. One quarter.
        v. One eighth.

    (c) If a satellite is changed to an orbit in which its period is octupled,[40] the radius of orbit will change by a factor of

        i. Eight.
        ii. Four.
        iii. Two.
        iv. One half.
        v. One quarter.
        vi. How the **** should I know?

[40] Dude: that means multiplied by 8.

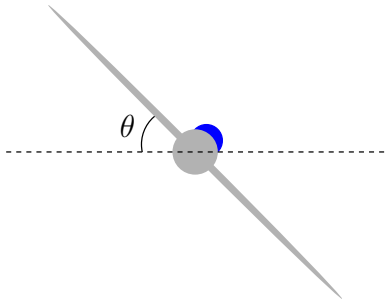


Figure 4.31: Problem 48

48. The wings of an airplane, for reasons that will be explained in §11.3, exert a force called the "lift" perpendicular to the plane of the airplane.[41] In straight, level flight, this lift is vertical and balances the weight of the plane.

    To make a turn, the plane "banks," as illustrated with a charming, almost childlike simplicity in fig. (4.31). If the plane has mass m and is moving at speed v, at what angle θ to the horizontal must the plane bank in order to make a horizontal turn of radius r? (Since this is not level flight, you should of course not assume that the lift force equals mg.)

[41] It was really tempting to phrase this as *perpendicular to the plane of the plane.* If only we were more of a bad-∗∗∗.

49. Suppose a half-chewed Gummy Bear that you drop down a wishing well experiences a frictional force proportional to its velocity.[42]

    (a) Show that if we take up to be the positive direction, the Gummy Bear's equation of motion is
    $$-mg - \beta v = ma$$
    where β is a positive constant.

    (b) If the Gummy Bear is dropped from rest, it will gain downward velocity because of the pull of gravity. But the faster it falls, the greater the air resistance. Eventually, air resistance will be great enough to balance the pull of gravity, and the Gummy Bear's velocity will level off at what is called the terminal velocity.

    Explain how you can cheat the Devil and obtain the result $-mg/\beta$ for the terminal velocity from $-mg - \beta v = ma$ without even solving this relation for the velocity as a function of time.

    (c) Assuming that the Gummy Bear is dropped from rest at time $t = 0$, solve its equation of motion to obtain its velocity as a function of time.

    (d) Make a rough sketch of v versus t (without using your calculator).

    (e) Verify that your result for $v(t)$ in #49c yields the terminal velocity you predicted in #49b. If you did not see how to do #49b, then you should go back and look at it again — using your result for the terminal velocity in #49b should make clear what you didn't see the first time around.

    (f) Explain how the dependence of your result for $v(t)$ on m and β makes sense physically.

    (g) Now solve the equation of motion to obtain velocity as a function of time if the Gummy Bear has a nonzero initial velocity v_0. You should find that this requires only relatively modest modifications to your work for the $v_0 = 0$ case.

    (h) You should find that your new result for $v(t)$ in #49g has an additional term in v_0 that dies away exponentially and that is therefore referred to as a transient. Make physical sense of this. That is, why should the effect of v_0 on your solution for $v(t)$ die away and thus become irrelevant at large t?

[42] Actually, for air resistance a frictional force proportional to v^2 would be much more realistic, but unfortunately that would make the integration much more difficult.

Also, it seems that in reality Gummy Bears actually have *negative* mass. At least, that's the only explanation we can think of for their being found on ceilings all over campus.

50. Okay, what the ∗∗∗∗: Suppose, much more realistically, that a half-chewed Gummy Bear you drop from rest down a wishing well experiences a frictional force proportional to the square of its velocity (that is, of magnitude βv^2, where β is a positive constant).

    (a) Determine the Gummy Bear's terminal velocity without even solving the equation of motion.

    (b) Assuming that the Gummy Bear is dropped at time $t = 0$, solve its equation of motion to obtain its velocity as a function of time.

    You will find it helpful to remember that

    $$\int \frac{dx}{a^2 - x^2} = \frac{1}{2a} \ln \frac{a + x}{a - x}$$

    (a result that you really should have been able to work out for yourself by the method of partial fractions) and that

    $$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

    (c) Make a rough sketch of v versus t (without using your calculator).

    (d) Explain how the dependence of this solution on (that is, the way it behaves because of the values of) m and β makes sense physically.

51. Suppose a mass m falling under the influence of the usual mg gravitational force also experiences a bizarre frictional force

$$F_f = \pm\kappa e^{-\beta v^2}$$

where κ and β are positive constants. The mass may be starting from rest or may have some initial downward velocity.

(Note that you never have to do an integration; this problem requires only careful quantitative reasoning. It may also help to remember that $e^{-\beta v^2} \leq 1$ always. Or it may not help. We'll see.)

(a) Describe the physical behavior of this frictional force. That is, describe how this frictional force behaves physically as a function of the parameters κ and β and the velocity v.

(b) Set up the equation of motion for the mass. This will involve determining which sign you want on the friction term.

(c) Derive a naive result for the mass's terminal velocity.

(d) Determine the restrictions, if any, on the ranges of values of κ and β for which your result for the terminal velocity is valid (it having already been given, of course, that κ and β are both positive).

(e) Show from the equation of motion that the mass experiences a downward acceleration when $\kappa < mg$ *or* it is falling at more than the terminal speed.

(f) Show from the equation of motion that the mass experiences an upward acceleration when $\kappa > mg$ *and* it is falling at less than the terminal speed.

(g) Make physical sense of the special case $\kappa = mg$ in the equation of motion. What is the "terminal" velocity in this case?

(h) Consider the case that the mass has some initial *upward* velocity. Describe the resulting motion.

(i) Describe all the ways in which the mass could end up reaching a finite terminal velocity.

(j) Explain whether this frictional force could realistically represent the friction experienced by a mass falling through the air or some other fluid.

(k) Explain whether this frictional force could realistically represent the friction experienced by a mass moving over a surface. (This is a question only about the frictional force, independent of there being any other applied force like mg.)

4.10 Sketchy Answers

(1a) $6.0\,\mathrm{m/s^2}$.

(1c) $20\,\mathrm{m/s^2}$.

(2) 8200 N.

(3) $\dfrac{m_b v_b^2}{2\ell}$.

(4a) 8.2 N.

(4b) 11 kg.

(4c) $0.16607 = \dfrac{1}{6.0215}$.

(8a) Like you don't know how to use a dictionary?

(8d) The set of answers you get should include

$$0, \quad \tfrac{2}{3}m_o g, \quad m_o g, \quad \tfrac{4}{3}m_o g, \quad m_y g, \quad \tfrac{2}{3}(m_o{+}m_y)g, \quad (m_o{+}m_y)g, \quad \tfrac{4}{3}(m_o{+}m_y)g$$

(15b) 140 N at 99°.

(15d) $4800\,\mathrm{m/s^2}$.

(15e) 480 m/s

(16a) 0.40.

(16b) 50 N.

(16c) $2.0\,\mathrm{m/s^2}$.

(16d) 500 N.

(16f) 750 N.

(16g) 470 N.

(17(a)i) $m'g(\mu_s \cos\theta + \sin\theta)$.

(17(a)ii) $m'g(\mu_k \cos\theta + \sin\theta)$.

(17(b)i) $\dfrac{m'g(\mu_s \cos\theta + \sin\theta)}{\cos\theta - \mu_s \sin\theta}$.

(17(b)ii) $\dfrac{m'g(\mu_k \cos\theta + \sin\theta)}{\cos\theta - \mu_k \sin\theta}$.

(19a) $\sqrt{\frac{10}{3} g\ell \sin\theta}$.

(21a) $a = \dfrac{mg}{m+M}, \quad T = \dfrac{mM}{m+M}\, g$.

(22a) $\dfrac{m_c}{m_a}\,\tan\theta$.

(23c) $a = \dfrac{m_r - m_o}{m_r + m_o}\,g, \quad T = \dfrac{2m_o m_r}{m_r + m_o}\,g.$

(24(a)i) $a = \dfrac{m\sin\theta - m'}{m + m'}\,g, \quad T = \dfrac{mm'}{m + m'}\,g(1 + \sin\theta).$

(24(b)i) $\dfrac{|m\sin\theta - m'|}{m\cos\theta}$.

(24(c)i) $a = \dfrac{m(\sin\theta + \mu_k\cos\theta) - m'}{m + m'}\,g$

$$T = \frac{mm'}{m + m'}\,g(1 + \sin\theta + \mu_k\cos\theta)$$

(25a) $\dfrac{F_{you}}{2m} - g$ and $\dfrac{F_{you}}{4m} - \frac{1}{2}g$.

(25b) $2m'g$.

(25(c)i) $\dfrac{F_{you}}{2m} - g, \quad \dfrac{F_{you}}{2m'} - g, \quad$ and $\frac{1}{4}F_{you}\left(\dfrac{1}{m} + \dfrac{1}{m'}\right) - g.$

(26a) $mg, \quad \dfrac{mg}{2\sin\theta}$.

(28a) $\dfrac{\mu_s mg}{\cos\theta - \mu_s\sin\theta}$.

(29) $mg\left(1 + \dfrac{\ell}{h_0 - h}\right)$ (or, with the values given, 40,000 N).

(30(a)i) 0.53 m/s.

(30(a)ii) 0.19.

(30(a)iv) 0.22.

(30(b)i) 29 m/s.

(30(b)ii) 2200.

(31a) $\dfrac{v^2}{gR} \pm 1$.

(34a) $mg\left(1 + \dfrac{v^2}{gR}\right)$.

(34b) $\sqrt{gR}$.

(35a) $T = \dfrac{mg}{\cos\theta}, \quad v = \sqrt{g\ell\sin\theta\tan\theta}.$

(36a) $\dfrac{mg\ell}{h}\left(\dfrac{\omega^2 h}{2g} \pm 1\right)$.

(37a) $T = m\left(\dfrac{v_A^2}{R} - g\sin\theta\right), \quad a_{tan} = g\cos\theta.$

(40a) $\dfrac{c^2 g}{\omega^2}$.

(40b) $\dfrac{\omega^6}{27c^2g^3}$.

(42) 1.5 hr.

(43) 3.59×10^7 m.

(46) 10.

(48) $\arctan\left(\dfrac{v^2}{gr}\right)$.

(49c) $-\dfrac{mg}{\beta}\left(1 - e^{-\frac{\beta}{m}t}\right)$.

(50a) $-\sqrt{\dfrac{mg}{\beta}}$.

(50b) $-\sqrt{\dfrac{mg}{\beta}}\,\tanh\left(\sqrt{\dfrac{\beta g}{m}}\,t\right)$.

Chapter 5

Work & Energy

> You could have a perpetual-motion machine that works for a while.
>
> Trey Kollmer [1]

Work is a bogus engineering concept; the really fundamental physical quantity is energy. What makes energy fundamental is that it's conserved: energy can change forms — from kinetic and the various kinds of potential energy into each other — , but the total energy, summed over all its forms, remains constant. Conservation of energy turns out to be a direct consequence of the symmetry of the universe under time translations: because the laws of physics are the same today as they were yesterday and will be tomorrow, there necessarily must be a conserved energy.

Probably you were blissfully unaware of this. And now maybe you're even thinking something along the lines of *"Whoa, dude! Whatta you been smokin'?"* or *"I wish my pizza would get here."* But bear with us for a moment. While it is very difficult to make it understandable at the level at which we are studying physics, it turns out that the fundamental principle in physics is symmetry: when we have figured out the overall symmetry of the universe, all of physics — everything that can exist and can happen physically — will follow logically and inevitably from this symmetry. It has been proved mathematically that for every physical symmetry there is a corresponding conserved quantity, and, contrary to what you might expect, it turns out that things that do not change are hugely important in physics: when a quantity remains constant, it puts constraints on what can exist and what can happen. Part of the symmetry of the universe is the invariance under time translations that we mentioned above, and we simply *define* energy

[1] A former student of the author. We pride ourselves on teaching students to think outside the box.

to be the corresponding conserved quantity that pops out when you do the math. The proof of this is presented in Appendix B. Although Appendix B is considerably beyond our present level — you would typically not get to this proof until graduate school — , we do want you to be aware that both the definition of energy and its conservation are direct mathematical consequences of one of the symmetries of nature. As we will see, one ramification of this is that what we define to be energy physically often does not gibe with the everyday uses of the word or with your intuitive notion of energy. In fact, there is not even a unique definition of energy: if energy E has a constant value, then so does $2E$ or $E + 42$ or $\arctan(\pi\sqrt{E})$ — the only important thing about energy is that it's conserved; its actual value is irrelevant.[2] So even if there were one definition of energy that completely encompassed your everyday, intuitive notions, we could, for the purposes of physics, instead choose another definition that had nothing in common with those notions.

Anyway, back to work. The only reason for introducing the notion of work at all is friction, and friction is, for reasons we'll discuss later, a bogus engineering force. If it were up to us, we wouldn't be covering either work or friction. But we don't make the rules. *Il faut cultiver notre jardin.*[3] And so ...

5.1 The Bogonics of Work & Power

At the level of Newtonian mechanics, the definitions of work and potential energy are obtained by integrating $\mathbf{F} = m\mathbf{a}$. We will first do the math to derive a result for work and then explain what this result means. Actually, we'll do the derivation two ways: once for the one-dimensional case and then again for the higher-dimensional case.

For one-dimensional motion along an x axis, we start with

$$F = ma = m\frac{dv}{dt}$$

With almost divine foresight, we will integrate both sides of this relation over the path followed by the body as it travels from an initial location x_i to a final location x_f:

$$\int_{x_i}^{x_f} F\,dx = \int_{x_i}^{x_f} m\frac{dv}{dt}\,dx$$

The integral on the left side cannot be carried out unless we know the force as a function of location, but with a bit of gerrymandering,[4] we can do the

[2]Though we should perhaps note that, while this is true in Newtonian physics, special relativity, and quantum field theory, in general relativity it remains difficult to this day even to determine a mathematically and physically viable definition of energy.

[3]From Voltaire's *Candide*.

[4]Isn't *gerrymandering* a great word?

integral on the right side: if we interchange the dv and the dx, we have

$$\begin{aligned}\int_{x_i}^{x_f} F\,dx &= \int_{x_i}^{x_f} m\frac{dv}{dt}\,dx \\ &= \int_{v_i}^{v_f} m\frac{dx}{dt}\,dv \\ &= \int_{v_i}^{v_f} mv\,dv \\ &= \tfrac{1}{2}mv^2\Big|_{v_i}^{v_f} \\ &= \tfrac{1}{2}mv_f^2 - \tfrac{1}{2}mv_i^2 \end{aligned} \tag{5.1}$$

where we have of course replaced the limits (x_i, x_f) for the dx integration to the corresponding limits (v_i, v_f) on the body's velocity when we are instead integrating by dv. This result leads us to make two definitions: first, we define

$$W = \int_{x_i}^{x_f} F\,dx \tag{5.2}$$

as the work *done by the force F on the body* as the body moves along its trajectory from x_i to x_f. Since we have derived this from $F = ma$, here the force F is the *net* force acting on the body, but definition (5.2) is applied in more general contexts to the work done by any force. Second, we define

$$K = \tfrac{1}{2}mv^2 \tag{5.3}$$

to be the kinetic energy of the body. With these definitions, eq. (5.1) becomes

$$W = K_f - K_i = \Delta K \tag{5.4}$$

Probably eqq. (5.2) and (5.3) don't seem in any way even remotely related to your intuitive notions of what constitutes work and energy. But remember that work and energy are defined purely mathematically in physics and that these definitions in fact in many ways do not gibe with your intuitive notions about these quantities. We can, however, make sense of eq. (5.4). Given that it is defined to be a form of energy, $\frac{1}{2}mv^2$ is called the *kinetic* energy because it depends only on the motion of the body: aside from the mass m, it depends only on the body's velocity v (as opposed to depending on the forces acting on the body or on other parameters not immediately related to the body's motion). Eq. (5.4) is telling us that the work done by the net force on the body determines the change in the body's kinetic energy — the same thing that $F = ma$ tells us, just phrased a bit differently: $F = ma$ told us that the net force on a body determines the body's motion in the sense of determining its acceleration; eq. (5.4) tells us that the work done by the net force on a body determines the body's motion in the sense of determining its

change in kinetic energy. $F = ma$ and eq. (5.4) are in fact the same relation: eq. (5.4) is just the integral of $F = ma$ and $F = ma$ is just the derivative of eq. (5.4).

Eq. (5.4) is often called the work-energy relation. We will be loathe to use this term for two reasons: first, work is, as we noted above, a bogus engineering concept, and second, the term *work-energy relation* makes it sound like eq. (5.4) is a fundamental relation when it certainly is not. But again, we don't make the rules, so you will be making some use of eq. (5.4), and when you do it is critical to bear in mind that while eq. (5.2) can be used to calculate the work done by any force, the W in eq. (5.4) is the work done by the *net* force.

The MKS unit of energy (any kind of energy) is the Joule (J). From eq. (5.4) we can see how the Joule is related to MKS units we have previously encountered: the dimensions of the left-hand side of eq. (5.4) being force times length and those of the right-hand side mass times the square of velocity, we have

$$1\ \mathrm{J} = 1\ \mathrm{N\,m} = 1\ \frac{\mathrm{kg\,m^2}}{\mathrm{sec^2}}$$

In higher dimensions, we must deal with $\mathbf{F} = m\mathbf{a}$ as a vector equation. Also, the infinitesimal displacements the body undergoes are no longer as simple as dx: the body is now undergoing vector displacements $d\mathbf{r}$ as it follows a higher-dimensional path from initial location $\mathbf{r}_i$ to final location $\mathbf{r}_f$. And instead of simply multiplying both sides of $F = ma$ by dx, we now want to take the dot product of both sides of $\mathbf{F} = m\mathbf{a}$ with $d\mathbf{r}$:

$$\mathbf{F} \cdot d\mathbf{r} = m\mathbf{a} \cdot d\mathbf{r} = m\frac{d\mathbf{v}}{dt} \cdot d\mathbf{r} = m\frac{d\mathbf{r}}{dt} \cdot d\mathbf{v} = m\mathbf{v} \cdot d\mathbf{v}$$

It would be nice if we could gerrymander the right-hand side into the same $\frac{1}{2}mv^2$ that we did in the one-dimensional case, and we can in fact do this if we note that, by the product rule,

$$d(v^2) = d(\mathbf{v} \cdot \mathbf{v}) = d\mathbf{v} \cdot \mathbf{v} + \mathbf{v} \cdot d\mathbf{v} = 2\mathbf{v} \cdot d\mathbf{v}$$

We thus have

$$\mathbf{F} \cdot d\mathbf{r} = \tfrac{1}{2}m\, d(v^2) = d(\tfrac{1}{2}mv^2)$$

When we integrate both sides from $\mathbf{r}_i$ to $\mathbf{r}_f$, with $\mathbf{v}_i$ and $\mathbf{v}_f$ as the corresponding limits on the velocity, this becomes

$$\int_{\mathbf{r}_i}^{\mathbf{r}_f} \mathbf{F} \cdot d\mathbf{r} = \int_{\mathbf{v}_i}^{\mathbf{v}_f} d(\tfrac{1}{2}mv^2) = \tfrac{1}{2}mv_f^2 - \tfrac{1}{2}mv_i^2$$

The right side of this relation is the same change of kinetic energy that we had in the one-dimensional case — except, of course, that v^2 is now the squared magnitude of the *vector* velocity (so that, in three dimensions,

$v^2 = v_x^2 + v_y^2 + v_z^2$). The left-hand side is what we define to be the work W done by the vector force F as the body moves along its trajectory from initial location $\mathbf{r}_i$ to final location $\mathbf{r}_f$:

$$W = \int_{\mathbf{r}_i}^{\mathbf{r}_f} \mathbf{F} \cdot d\mathbf{r} \tag{5.5}$$

The integration in eq. (5.5) is a line integral of the type discussed in §2.1. We will postpone saying more about the significance of this until we discuss potential energy in §5.2.

In the meantime, a frequently occurring special case of eq. (5.5) is the work done by a constant force. When $\mathbf{F}$ is constant, we can pull it outside of the integration, to obtain

$$W = \int_{\mathbf{r}_i}^{\mathbf{r}_f} \mathbf{F} \cdot d\mathbf{r} = \mathbf{F} \cdot \int_{\mathbf{r}_i}^{\mathbf{r}_f} d\mathbf{r} = \mathbf{F} \cdot (\mathbf{r}_f - \mathbf{r}_i) = \mathbf{F} \cdot \Delta\mathbf{r}$$

That is, for a constant force the work done is simply the dot product of the force with the *overall* displacement $\Delta\mathbf{r} = \mathbf{r}_f - \mathbf{r}_i$. If we use the geometric definition (1.2) of the dot product, this becomes

$$W = F\,|\Delta\mathbf{r}|\cos\theta$$

where θ is the angle between the force $\mathbf{F}$ and the overall displacement $\Delta\mathbf{r}$. With the somewhat simpler conventional notation s for the magnitude of this overall displacement (see fig. (5.1)), we arrive at

$$W = Fs\cos\theta \tag{5.6}$$

The $\cos\theta$ in eq. (5.6) (and in fact the $\mathbf{F} \cdot d\mathbf{r}$ in eq. (5.5)) is telling us that a force does work on a body only to the extent that it acts along the body's line of motion; to the extent that a force is at a right angle to the body's motion, it does no work. This in fact makes sense according to eq. (5.4), $W = \Delta K$: to the extent that a force acts along a body's direction of motion, it is causing the body to speed up or slow down and thus increasing or decreasing the body's kinetic energy; to the extent that a force is perpendicular to a body's direction of motion, it is changing only the body's direction of motion — there is no change in speed and hence no change in kinetic energy.

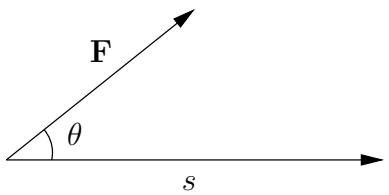


Figure 5.1: Case of a Constant Force

Note that a constant force must be constant as a vector, that is, constant in both magnitude and direction. The gravitational force mg is an example of a constant force. Friction, however, is not: even if the magnitude of a frictional force is constant, its direction at any moment is opposite to the direction in which the object experiencing the frictional force is moving at that moment. The work done by a frictional force therefore depends on the path taken and not just on the overall displacement, so that eq. (5.6) does not hold for frictional forces.

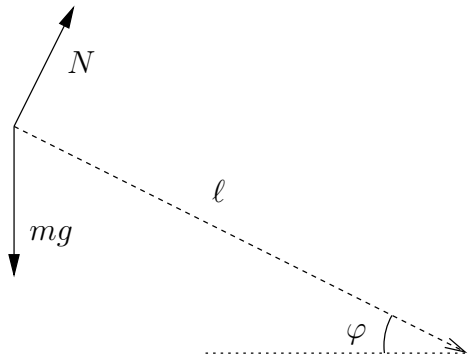


Figure 5.2: Butt-Sledding Down a Hill

As an example of work done by constant forces, suppose that one fine snowy day you (mass m) go butt-sledding down a hill of length ℓ inclined at a conveniently uniform angle φ to the horizontal. Under the assumption that the hill is icy enough that friction can be neglected and that therefore the normal and gravity are the only forces acting on you, your force diagram is as shown in fig. (5.2). Both N and mg are constant forces, so we can use eq. (5.6) to calculate the work they do on you. Since the normal is at a right angle to your displacement down the incline, the work done by the normal force is

$$W_N = N\ell\cos\theta_N = N\ell\cos\frac{\pi}{2} = 0$$

This result is quite general: when a body is moving along a stationary surface, *the normal force does no work* because it is always at a right angle to the body's direction of motion.[5] Since the angle between your displacement down the incline and mg is $\frac{\pi}{2} - \varphi$, the work done by gravity is

$$W_{mg} = mg\ell\cos\theta_{mg} = mg\ell\cos(\frac{\pi}{2} - \varphi) = mg\ell\sin\varphi$$

[5]Note that this does not, however, mean that the work done by normal forces vanishes under all circumstances. When you are in a moving elevator, for example, both your displacement and the normal force between you and the floor are vertical, and consequently work done by the normal force is nonzero.

The net work done on you (which is the same thing as the work done on you by the net force) is thus

$$W_N + W_{mg} = 0 + mg\ell \sin\varphi = mg\ell \sin\varphi$$

If you had a running start at speed v_0, eq. (5.4) would therefore give

$$\begin{aligned} W &= \tfrac{1}{2}mv_f^2 - \tfrac{1}{2}mv_i^2 \\ mg\ell \sin\varphi &= \tfrac{1}{2}m(v_f^2 - v_0^2) \\ v_f &= \sqrt{v_0^2 + 2g\ell \sin\varphi} \end{aligned}$$

Gravity does a positive amount of work on you, with the result your kinetic energy and hence speed increase.

This result for your velocity v_f at the bottom of the hill is of course exactly the same as we would have obtained from constant acceleration equations with the familiar acceleration $a = g\sin\varphi$ along an incline:

$$\begin{aligned} v_f^2 - v_0^2 &= 2a(x - x_0) = 2g\sin\varphi\,(\ell) \\ v_f &= \sqrt{v_0^2 + 2g\ell \sin\varphi} \quad \checkmark \end{aligned}$$

If, however, the inclination of the hill were not uniform, so that instead of being straight its surface were some sort of curve, then your acceleration down the hill would not be constant and we could not use constant-acceleration relations to solve for v_f. But as long as your overall displacement between your starting point at the top of the hill and your endpoint at the bottom of the hill remained the same, we would get the same results for the work done by the normal and gravity and thus the same result for v_f.

Power P is simply the rate of doing work:

$$P = \frac{dW}{dt} \tag{5.7}$$

From eq. (5.5), the work dW done as the body undergoes the infinitesimal displacement $d\mathbf{r}$ is $dW = \mathbf{F} \cdot d\mathbf{r}$. Using this in eq. (5.7), we have

$$P = \frac{dW}{dt} = \frac{\mathbf{F} \cdot d\mathbf{r}}{dt} = \mathbf{F} \cdot \frac{d\mathbf{r}}{dt} = \mathbf{F} \cdot \mathbf{v}$$

so that

$$P = \mathbf{F} \cdot \mathbf{v} = Fv\cos\theta \tag{5.8}$$

where θ is the angle between the force F and the body's velocity v.

The MKS units of power are Watts (W): 1 W = 1 J/s. Thus a 60 W incandescent light bulb consumes 60 J of electrical energy each second (only a small fraction of which is converted into light, the rest producing waste

heat). One horse power (hp) is about 745.7 W, though it's anybody's guess how they managed to come up with such a precise definition when there is so much individual variation between horses. It's also a mystery why there isn't a unit called "people power" (pp) when it seems like such an obvious and natural unit to define. If there were such a unit, some sort of Kyoto-like protocol might define it to be about 200 W, the level of aerobic output sustainable for long periods by someone in good shape. But then the United States would probably wouldn't agree to the whole thing, arguing for a lower figure more consistent with the output of the typical suburban-assault-vehicle-driving American couch potato.

5.2 Potential Energy & Energy Conservation

The really fundamental physical quantity is not work but rather energy, and energy turns out to be of two types: kinetic (the same as in the preceding section) and potential.

The derivations of the relations for energy and potential energy are very similar to those for work, but with some important differences. In the one-dimensional case, we again start from $F = ma$, but this time we move everything to the same side of the equation:

$$0 = ma - F = m\frac{dv}{dt} - F$$

Integrating both sides by dx, we obtain

$$\begin{aligned} 0 &= \int_{x_i}^{x_f} m\frac{dv}{dt}\,dx - \int_{x_i}^{x_f} F\,dx \\ &= \int_{v_i}^{v_f} m\frac{dx}{dt}\,dv - \int_{x_i}^{x_f} F\,dx \\ &= \int_{v_i}^{v_f} mv\,dv - \int_{x_i}^{x_f} F\,dx \\ &= \tfrac{1}{2}mv_f^2 - \tfrac{1}{2}mv_i^2 - \int_{x_i}^{x_f} F\,dx \\ &= \Delta K - \int_{x_i}^{x_f} F\,dx \end{aligned} \tag{5.9}$$

where we are using the same definition $K = \frac{1}{2}mv^2$ for kinetic energy as before. The integral of the force F is also the same as we had before, but now, although it might seem like an odd thing to do, we will break it up by integrating first from x_i to some arbitrary point x_0 and then from x_0 to x_f:

$$-\int_{x_i}^{x_f} F\,dx = -\left(\int_{x_i}^{x_0} F\,dx + \int_{x_0}^{x_f} F\,dx\right)$$

If we flip the limits of integration on the first integral on the right-hand side, this becomes

$$\begin{aligned} -\int_{x_i}^{x_f} F\,dx &= -\left(-\int_{x_0}^{x_i} F\,dx + \int_{x_0}^{x_f} F\,dx\right) \\ &= -\left(-\int_{x_0}^{x_i} F\,dx\right) + \left(-\int_{x_0}^{x_f} F\,dx\right) \\ &= -U(x_i) + U(x_f) \\ &= \Delta U \end{aligned}$$

where we have defined the potential energy $U(x)$ at location x to be

$$U(x) = -\int_{x_0}^{x} F\,dx \tag{5.10}$$

With this definition of potential energy, eq. (5.9) becomes

$$0 = \Delta K + \Delta U = \Delta(K + U)$$

Thus the change in the quantity $K + U$ is zero, which means that the value of $K + U$ is constant. We define this sum $K + U$ to be the total energy E:

$$E = K + U \tag{5.11}$$

Although we have derived it from $F = ma$ rather than from symmetry principles, the energy E of eq. (5.11) is the conserved quantity corresponding to the invariance of physics under time translations and eq. (5.11) is the law of conservation of energy. Again, the important thing about this energy E is that its value is constant. The kinetic energy K and potential energy U are, however, not separately conserved — they each vary over the trajectory of the body — , but since the sum $K + U$ remains constant, as K increases, there is a commensurate decrease in U and vice versa. In other words, it is as though over the trajectory of the body there is a continual exchange between kinetic and potential energies. Hence the name *potential energy*: the potential energy U is like a stored energy that has the potential to produce motion by increasing the kinetic energy K. Conversely, when a body slows down, its kinetic energy is being converted into potential energy.[6] Thus one commonly speaks of potential energy being *liberated* or *released* as it is converted to kinetic energy, or of the kinetic energy of a body being *stored* as some form of potential energy.

The point x_0 in eq. (5.10) is, as we said, arbitrary: we can choose x_0 to be any location we want; x_0 *does not even have to be on the actual trajectory of the body.* From eq. (5.10), we see that

$$U(x_0) = -\int_{x_0}^{x_0} F\,dx = 0$$

[6] With the exception of the perverse case of friction, which we will deal with shortly.

That is, the value of the potential energy U is always zero at x_0. In other words, choosing x_0 is equivalent to deciding where to put the zero of the potential energy. In practice, you usually choose an x_0 that makes the expression for the potential energy U as simple as possible. (This will become more clear when we work out results for the potential energies corresponding to various forces in the next section.)

In higher dimensions, eq. (5.10) becomes

$$U(\mathbf{r}) = -\int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{F} \cdot d\mathbf{r} \tag{5.12}$$

where the location $\mathbf{r}$ is arbitrary. In this higher-dimensional case there is, however, a mathematical complication. It turns out that the definition (5.12) yields a monovalued function $U(\mathbf{r})$ only for certain forces F, and a function that is not monovalued makes no sense physically: how can a physical quantity have more than one value at a given place and time? To see the nature of the restrictions on the force F, consider the case $\mathbf{r} = \mathbf{r}_0$ in eq. (5.12):

$$U(\mathbf{r}_0) = -\int_{\mathbf{r}_0}^{\mathbf{r}_0} \mathbf{F} \cdot d\mathbf{r}$$

We of course want this integral to be zero, because its upper and lower limits are the same. But because we are now in two or more dimensions, there are an infinite number of paths that will take us from $\mathbf{r}_0$ back to $\mathbf{r}_0$. To get uniquely zero for the above integral, we must get zero for it over all possible paths. In §2.6.1, it was shown that the vanishing of the line integral of a vector function around an arbitrary closed loop is equivalent to the vanishing of the curl of that vector function. In order for the result of the integral for U to be unique, the force must therefore have zero curl:

$$\nabla \times \mathbf{F} = 0$$

Forces satisfying this condition are said to be conservative.[7] All truly physical forces turn out to be conservative, and it is therefore possible to define potential energies corresponding to them.

Friction, however, is not a conservative force. Consider, for example, the bogus kinetic frictional force $F_f = \mu_k N$ for a box sliding on a floor: this frictional force is always opposite in direction to the displacement $d\mathbf{r}$ of the box, so that

$$\mathbf{F}_f \cdot d\mathbf{r} = F_f \, dr \cos \pi = -F_f \, dr$$

Since $F_f = \mu_k N$ is constant in magnitude, when we integrate $\mathbf{F}_f \cdot d\mathbf{r}$, we will get simply

$$-\int F_f \, dr = -F_f \int dr = -F_f \times (\text{distance moved})$$

[7] The more general mathematical term for a vector field of vanishing curl is irrotational.

The result we get when we evaluate this integral therefore depends on the path along which we move the box: different paths will cover different distances. Since the value of the integral is not unique, it is not possible to define a potential energy for frictional forces: to deal with friction, you have to use work instead of potential energy, because the line integral for work is evaluated only along the actual trajectory of the body and therefore yields a unique result.[8] This is the only reason that we bother with work at all.

On a microscopic level, friction is just a manifestation of electromagnetic forces between the myriad pairs of molecules in two bodies. There is therefore no such thing as a frictional force per se — looked at on a small enough scale, what appears to be a distinct frictional force is just a special case of the electromagnetic force. It is just that on a macroscopic level we neither see nor are able to account for the myriad microscopic interactions, so we attribute their effect to a bogus force that we call friction. It's just like macro- and microeconomics: all economics is really just people selling potatoes and buying shoes, but at the macroeconomic level, it's impossible to see how these myriad little transactions bring about the overall economy, so economists just make up a bunch of bogus concepts and pretend that they are real and have meaning.

Anyway, energy is really still conserved even when frictional forces are acting; it is just that what had been macroscopic kinetic energy visible in the motions of bodies is converted into the invisibly microscopic form of energy we call heat — the kinetic energy of random molecular motion.

Again, what matters about the energy $E = K+U$ is not its value, but that its value remains constant. We can change the value of E arbitrarily just by changing the point x_0 in eq. (5.10) or $\mathbf{r}_0$ in eq. (5.12) from which we measure our potential energy U: this will have the effect of everywhere shifting the value of U, and hence of E, up or down by some fixed amount. The values of the potential energy U and total energy E at any particular point, being arbitrary, therefore have no physical meaning;[9] all that matters is how the potential energy *changes* as we go from one point to another: *bodies always tend to move to lower potential energy.* This happens precisely because of the way potential energy is defined in eq. (5.10): since the potential energy

[8] It is, however, still possible to use $E = K + U$ rather than $W = \Delta K$ if we modify our energy conservation relation to take into account the energy "lost" to friction, as we (or rather, you) will do in problem #28.

[9] By that curious irony that seems to govern the universe, when you get to quantum field theory, it turns out that the value of the potential energy is in fact the only thing that matters. The electromagnetic potential, for example, turns out to be the photon (the particle of light): the value of the electromagnetic potential at each point in spacetime tells you how many photons there are at that place and time.

U is the negative integral of the force F, F is the negative derivative of U:

$$F = -\frac{dU}{dx} \tag{5.13}$$

Bodies will, of course, tend to move in the direction of the force exerted on them, and this relation tells us that the force is in the direction of $-dU$, that is, in the direction of negative change (decrease) in potential energy.

To see how this works out in higher dimensions, consider the single infinitesimal contribution to the integration of eq. (5.12) from a single infinitesimal displacement $d\mathbf{r}$:

$$dU = -\mathbf{F} \cdot d\mathbf{r}$$

According to eq. (2.5), the change dU in a function $U(\mathbf{r})$ can in general be written in terms of its gradient:

$$dU = \nabla U \cdot d\mathbf{r}$$

Putting the above two relations together, we have

$$\nabla U \cdot d\mathbf{r} = -\mathbf{F} \cdot d\mathbf{r}$$

Since this relation must hold for *all possible* $d\mathbf{r}$, we must have [10]

$$\mathbf{F} = -\nabla U \tag{5.14}$$

On p.95 we established that ∇U is always in the direction of the steepest rate of increase of U. Eq. (5.14) is thus telling us that the force is not only toward lower potential energy but, more precisely, is in the direction of the steepest descent of U.

5.3 A Practical Example of Energy Conservation

In addition to being one of the most fundamental laws governing the physical universe, conservation of energy is also useful at a more pedestrian level for solving certain kinds of physical problems: since the value of $K+U$ is constant throughout the motion of a body, we can in particular set

$$K_1 + U_1 = K_2 + U_2$$

for any two points 1 and 2 on the trajectory. If we have enough information to evaluate the sum $K + U$ at point 1, often we can then determine the value

[10] The logic being that if $\mathbf{F}$ were not equal to $-\nabla U$, we would be able to find some $d\mathbf{r}$ for which $\nabla U \cdot d\mathbf{r}$ would not be equal to $-\mathbf{F} \cdot d\mathbf{r}$.

of the velocity (or whatever other unknown physical quantity is of interest to us) at point 2.

Suppose, for example, that you are trying to place a large plastic Santa on a snowy rooftop when you slip and slide from rest, with negligible friction, down the roof, off the edge, and into a nasty clump of thorny rose bushes a vertical distance of 12 m below the point where you began your descent. As we will see in the next section, the potential energy corresponding to the gravitational force mg is $U = mgy$, where the positive y direction is up. If we apply energy conservation with the first point being the start of the slide (denoted by a subscript s) and the second point the impact (denoted by a subscript i), then $K_1 + U_1 = K_2 + U_2$ becomes

$$\tfrac{1}{2}mv_s^2 + mgy_s = \tfrac{1}{2}mv_i^2 + mgy_i$$

Your mass m will cancel out, and, since you start from rest, $v_s = 0$. Thus

$$gy_s = \tfrac{1}{2}v_i^2 + gy_i$$

or

$$v_i = \sqrt{2g(y_s - y_i)} \tag{5.15}$$

In keeping with general arbitrariness of the value of the potential energy, we can put the origin of our y axis anywhere we want: since your starting point is 12 m above your final point, any choice where y_s is 12 m greater than y_i will do: we could choose $y_s = 12$ m and $y_i = 0$, or $y_s = 0$ and $y_i = -12$ m, etc. And from our solution (5.15) it is clear that all of these infinitely many choices of origin will yield the same v_i, since in every case $y_s - y_i = 12$ m. Thus we have

$$v_i = \sqrt{2(9.8)(12)} = 15 \text{ m/s}$$

or about 34 mph. Ouch.

Note that, as in our previous $W = \Delta K$ example, when we do this problem by conservation of energy, we need know only the speed at which you started to slip and the vertical distance through which you descended; we don't need to know the shape or extent of the roof or the angle at which you leave the edge and go airborne. Conservation of energy is a powerful method that often makes the problem much simpler than other methods. But there is a price: when we use conservation of energy (or $W = \Delta K$), we can calculate only your speed and not the direction in which you are moving, how far you have traveled horizontally, or the like.

When setting up energy conservation, you always want to ask yourself what forces are affecting the motion, so that you can include the corresponding potential energies in $K + U = K + U$. If, for example, you are bungee jumping, the forces affecting your motion are gravity and the force exerted

by the bungee cord, and you should therefore include the potential energies corresponding to both of these forces in your energy equation:

$$K_1 + U_{1,\,bungee} + U_{1,\,grav} = K_2 + U_{2,\,bungee} + U_{2,\,grav}$$

It may have seemed odd to you that in the above example we omitted any potential energy corresponding to the normal force, since there certainly would be a considerable normal force acting on you while you were sliding on the roof. In some ways the normal force, like friction, is bogus: it is a macroscopic manifestation of what on a microscopic level turn out to be electromagnetic forces. It would not even be possible to define a potential energy corresponding to a normal force, not least because a normal force isn't defined at points off of the trajectory of the body. But in §5.1 we saw that when a body is moving over a surface the normal force doesn't do any work, and since the normal force therefore doesn't change the body's kinetic energy, it may be taken to have no potential energy associated with it. This keeps life simple for bodies moving over surfaces.

Unfortunately, as also noted in §5.1, there are circumstances under which the normal force will do work. When you are in a moving elevator, for example, both your displacement and the normal force between you and the floor are vertical, and consequently the work done by the normal force is nonzero. In such situations it is, however, still possible to use $E = K + U$ rather than $W = \Delta K$ if we modify our energy conservation relation to take into account the work done by the normal force, as you will do in problem #29.

5.4 Results for Potential Energy

With definitions (5.10) and (5.12), we can obtain results for the potential energies corresponding to some common forces.

Gravity Near the Earth's Surface

If, as usual, we take the positive y direction to be vertically upward, then the force of gravity is $-mg\,\hat{\mathbf{y}}$. In Cartesian coordinates, the general expression for a three-dimensional vector displacement $d\mathbf{r}$ is

$$d\mathbf{r} = d(x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} + z\,\hat{\mathbf{z}}) = dx\,\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}} + dz\,\hat{\mathbf{z}}$$

Using all this in eq. (5.12), we have

$$U_{grav} = -\int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{F} \cdot d\mathbf{r}$$

$$
\begin{aligned}
&= -\int_{\mathbf{r}_0}^{\mathbf{r}} -mg\,\hat{\mathbf{y}}\cdot(dx\,\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}} + dz\,\hat{\mathbf{z}}) \\
&= -\int_{\mathbf{r}_0}^{\mathbf{r}} -mg\,(dx\,\hat{\mathbf{y}}\cdot\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}}\cdot\hat{\mathbf{y}} + dz\,\hat{\mathbf{y}}\cdot\hat{\mathbf{z}}) \\
&= -\int_{\mathbf{r}_0}^{\mathbf{r}} -mg\,\Big(dx(0) + dy(1) + dz(0)\Big) \\
&= -\int_{y_0}^{y} -mg\,dy \\
&= mg(y - y_0)
\end{aligned}
$$

We are free to choose any y_0 we find convenient, so we will use $y_0 = 0$ in order to make the result for the potential energy as simple as possible. We then have

$$U_{grav} = mgy \tag{5.16}$$

When using eq. (5.16), remember that the positive y direction has already been chosen to be upward. You are, however, free to place the origin of your y axis anywhere you want.

Newtonian Gravity

The math of this case is a little more involved. If we set up our coordinate system so that mass m_1 is at the origin, the Newtonian gravitational force felt by m_2 is toward m_1, that is, radially inward toward the origin. We therefore want to use the expression (2.4) for $d\mathbf{r}$ in spherical coordinates when we set up eq. (5.12):

$$d\mathbf{r} = dr\,\hat{\mathbf{r}} + r\,d\theta\,\widehat{\boldsymbol{\theta}} + r\sin\theta\,d\varphi\,\widehat{\boldsymbol{\varphi}}$$

We also want to express the force (which, remember, we are taking specifically to be the force exerted on m_2 and which is radially inward toward the origin) as

$$\mathbf{F} = -\frac{Gm_1m_2}{r^2}\,\hat{\mathbf{r}}$$

Thus eq. (5.12) becomes

$$
\begin{aligned}
U_{grav} &= -\int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{F}\cdot d\mathbf{r} \\
&= -\int_{\mathbf{r}_0}^{\mathbf{r}} -\frac{Gm_1m_2}{r^2}\,\hat{\mathbf{r}}\cdot(dr\,\hat{\mathbf{r}} + r\,d\theta\,\widehat{\boldsymbol{\theta}} + r\sin\theta\,d\varphi\,\widehat{\boldsymbol{\varphi}}) \\
&= -\int_{\mathbf{r}_0}^{\mathbf{r}} -\frac{Gm_1m_2}{r^2}\,(dr\,\hat{\mathbf{r}}\cdot\hat{\mathbf{r}} + r\,d\theta\,\hat{\mathbf{r}}\cdot\widehat{\boldsymbol{\theta}} + r\sin\theta\,d\varphi\,\hat{\mathbf{r}}\cdot\widehat{\boldsymbol{\varphi}}) \\
&= -\int_{\mathbf{r}_0}^{\mathbf{r}} -\frac{Gm_1m_2}{r^2}\,\Big(dr\,(1) + +r\,d\theta\,(0) + r\sin\theta\,d\varphi\,(0)\Big)
\end{aligned}
$$

$$= \int_{r_0}^{r} \frac{Gm_1m_2}{r^2}\,dr$$
$$= -\frac{Gm_1m_2}{r} + \frac{Gm_1m_2}{r_0}$$

The simplest choice this time is $r_0 = \infty$, so that we end up with

$$U_{grav} = -\frac{Gm_1m_2}{r} \tag{5.17}$$

Springs

Every spring has a certain natural, relaxed length; if one end of the spring is anchored (say, to a wall) and an object is attached to the other end, the spring will not exert any force on the object when it is at this relaxed length. When the spring is distended or compressed, it exerts a force that always pushes or pulls the object back toward this equilibrium position. The force law for real spring forces can be quite complicated; we will concern ourselves only with an idealized sort of spring known as a harmonic spring, for which the spring force is proportional to the distance by which the spring is stretched or compressed from its equilibrium length.

If we lay our x axis along the spring with $x = 0$ at the equilibrium point, then x does in fact measure how far the spring has been stretched or compressed from its equilibrium length, and the force law may be expressed analytically as

$$F_{spring} = -kx\ , \tag{5.18}$$

where k, the constant of proportionality, is called the spring constant of the spring. The negative sign in this relation corresponds to the spring force being a restoring force, that is, being always back toward equilibrium: in whichever direction, positive or negative x, we displace the end of the spring from its equilibrium position, the spring force is always the "other" way, that is, back toward the equilibrium position.

Some books omit the negative sign in the spring force law. That's kid stuff. Many books also refer to this force law as "Hooke's law." That's a far worse offense: it makes it sound as though $F = -kx$ were some sort of fundamental physical law, like conservation of energy. The real truth is this:

- When applied to springs, harmonic force laws like eq. (5.18) are important not because they are fundamental physical laws — they most certainly are not — , but because they are good engineering approximations for the spring forces.

- Harmonic force laws like eq. (5.18) are more important because they are also a good approximation for the forces exerted by numerous and diverse systems that are in some sense in equilibrium and that are subjected to reasonable disturbances. For example, your house will oscillate in a moderate earthquake just like a mass oscillates on the end of a spring.[11]

- Harmonic force laws like eq. (5.18) are still more important because, in an abstract mathematical sense, everything that exists is like a harmonic spring. When you do quantum field theory, there is a field corresponding to every kind of matter that exists in nature (photons, electrons and positrons, etc.). The value of this field tells you how many particles of that kind of matter you have at each point in spacetime and is determined by an equation virtually identical to the equation governing the motion of a mass on the end of a harmonic spring. Interactions between particles, which account for everything that can happen in the physical universe, are like interactions of neighboring springs. In string theory, the picture is very similar, except that the oscillations corresponding to particles are the vibrations of a one-dimensional object (the string) rather than being point-like. This fundamental similarity between spring oscillations and the relations governing quantum field and string theory is the real reason physics texts always make such a big deal out of springs.

Anyway, back to potential energy: The force law (5.18) in eq. (5.10) gives us

$$U_{spring} = -\int_{x_0}^{x} F\,dx = -\int_{x_0}^{x} -kx\,dx = k\int_{x_0}^{x} x\,dx = k(\tfrac{1}{2}x^2 - \tfrac{1}{2}x_0^2)$$

The convenient choice for x_0 is $x_0 = 0$. We then have

$$U_{spring} = \tfrac{1}{2}kx^2 \tag{5.19}$$

When using eqq. (5.18) and (5.19), you can make distention the positive x and compression the negative x direction or vice versa — the choice is arbitrary. But once you have made a choice, you of course need to be consistent about it.

5.5 How to Beat a Dead Horse

We've covered all the energy stuff and said a lot about why it is fundamental and important in principle. Now let's be sure that you are clear on what

[11] More precisely, your house will respond like a complex, three-dimensional system of coupled springs.

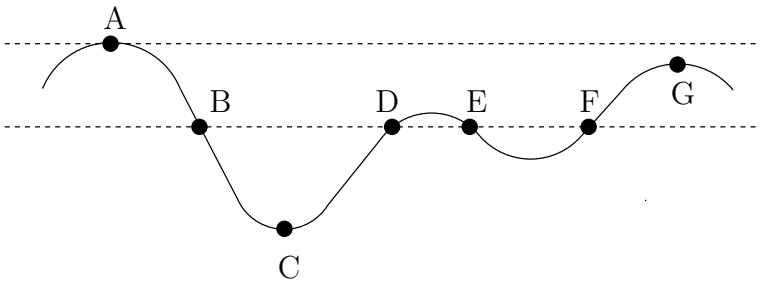


Figure 5.3: A Roller Coaster

conservation of energy means in practice in its more mundane, everyday applications.

Consider the roller coaster shown in fig. (5.3), along which is rolling and coasting a cart of total mass m (including the riders). If friction is negligible, the only forces affecting the motion of the cart are gravity and a normal force exerted on the cart by the track. Since there is no potential energy associated with the normal force, the only potential in our conservation of energy relation will be the gravitational potential mgy:

$$\tfrac{1}{2}mv_1^2 + mgy_1 = \tfrac{1}{2}mv_2^2 + mgy_2$$

for any two points 1 and 2 in the motion. This relation is telling us that, since the sum of the two terms $\frac{1}{2}mv^2$ and mgy is constant throughout the motion, the larger y, the smaller v and vice versa: as the cart descends, gravitational potential energy is released and converted into kinetic energy, with the result that the cart gains speed; as the cart ascends, its kinetic energy is being converted into and stored as gravitational potential energy, with the result that the cart loses speed. And at any two points where the cart is at the same height y, it will have the same speed v. Thus in fig. (5.3), the cart will be at its slowest at its highest point (A) and at its fastest at its lowest point (C). Since point G is slightly lower than point A, the cart will go over peak G at a slightly higher speed than it went over peak A. And at points D, E, and F the cart will have the same speed that it did at point B.

In real roller coasters, a sort of winch pulls you up to and just over the top of the first peak, after which your motion is (modulo friction, which the designers do their best to minimize) entirely determined by gravitational potential energy. This is why the first peak is always high, both in absolute terms and relative to all the other peaks on the ride: you are hauled over this first peak only very slowly, and virtually all of your subsequent motion comes from the release of your gravitational potential energy at this highest point.

Consider also bungee jumping. As we mentioned earlier, the forces affecting your motion, if we neglect air resistance, are gravity and the force

exerted on you by the bungee cord. If the bungee cord acts like a harmonic spring, then our conservation-of-energy relation will be

$$K_1 + U_{1,\,bungee} + U_{1,\,grav} = K_2 + U_{2,\,bungee} + U_{2,\,grav}$$
$$\tfrac{1}{2}mv_1^2 + \tfrac{1}{2}kx_1^2 + mgy_1 = \tfrac{1}{2}mv_2^2 + \tfrac{1}{2}kx_2^2 + mgy_2$$

Now there are three forms among which your energy exchanges itself: kinetic energy, bungee potential, and gravitational potential. As you fall, at first the bungee cord is slack and your gravitational potential is being converted entirely into kinetic energy, with the result that you speed up (rather dramatically, from a human perspective). At some point, the bungee begins to stretch. Since you are still descending, you are still releasing gravitational potential, but now this released gravitational potential is no longer being converted entirely into kinetic energy; some of it is being stored in the bungee potential. Since the bungee potential is quadratic in the distance x that the bungee is stretched but the gravitational potential you are releasing is only linear in the distance you are falling, the bungee is able to soak up not only the additional gravitational potential you are releasing, but also the kinetic energy you had acquired before the bungee began to stretch. Thus you slow down and eventually come to rest. At this lowest point, overall energy has been converted from gravitational potential to bungee potential. If there were no friction, energy conservation dictates that the whole process would then reverse itself and you would return to the platform, coming to rest just as you reached it. Although that would be really kind of cool, in reality you don't go anywhere near that high on the bounce because there is in fact considerable friction, both within the bungee cord and also between you and the air.

If the bungee cord is too long,[12] all of your kinetic energy seems to suddenly disappear when you crash into the pavement, but this loss of energy is only apparent: your macroscopic kinetic energy is converted on impact into the microscopic molecular motion that constitutes heat (less, of course, any energy used for breaking bones, etc.).

[12]This has been the occasion of at least one Darwin Award.

5.6 Problems

1. In a kind of metaphor for life, you push as hard as you can against a brick wall until you pass out from exhaustion.

   (a) Have you done any work physiologically? If so, where does this energy come from?

   (b) Have you done any work physically? Explain. If so, where does this energy come from?

2. In a friendly pickup soccer game, you take out an opponent with a vicious slide tackle from behind. If you (mass m) go from speed V to rest as you slide a distance ℓ along the field, how much work is done by the frictional force that brings you to rest?

3. Using the definition of work, show that the work done by kinetic friction is always negative. Explain how this makes sense in terms of the "work-energy relation" $W = \Delta K$.

4. A table of bullet masses and muzzle velocities is given below for a few popular calibers. Which is most lethal and why?

Caliber	Mass (g)	Velocity (m/s)
.357 Magnum	10.2	440
.44 Magnum	15.5	440
.223 (M16)	3.6	970

5. (The Adventures of Spot.) See Spot run! Run, Spot, run! "Upon my word," said farmer Brown. Ah, those were the days.

   But enough reminiscing; down to business. Suppose you take your dog Spot (mass m) out for a walk and that when you are a distance ℓ from home, Spot keels over and dies.[13] You drag Spot home over level ground by pulling on the leash at an angle of θ to the horizontal. There is friction between Spot and the ground, and you find that you have to pull with a constant force F_{you} to keep Spot moving at a constant speed V.

   (a) Over the distance ℓ of the trip home
      i. How much work do you do?
      ii. How much work is done by gravity?
      iii. How much work is done by friction?
      iv. How much work is done by the normal force?

   (b) Show that
   $$W_{you} + W_{gravity} + W_{friction} + W_{normal} = 0$$
   For what physical reason do you expect this to be so? (Note that we are looking for a *physical* reason. A savage beating with the foam noodle awaits anyone who replies "because they cancel each other out.")

   (c) How much power are you putting into dragging Spot? That is, at what rate are you doing work on him?

   (d) You should have found that while the power you put into dragging Spot depends on the speed V at which you are pulling him, the work you do dragging him home does not. Explain how this nonetheless makes sense.

   (e) Will this problem ever end?

6. (The Return of Spot.) Suppose Spot (still mass m) is not on level ground, but on a decline at angle θ to the horizontal, and that you have to pull with a constant horizontal force F_{you} to keep Spot moving at a constant speed V. That is, the distance ℓ that you have to drag Spot along the sidewalk to get home is tilted downward at θ to the horizontal, and the leash is horizontal as you pull on it.

   Over the distance ℓ of the trip home

   (a) How much work do you do?

   (b) How much work is done by gravity?

   (c) How much work is done by friction?

   (d) How much work is done by the normal force?

[13] Don't get the wrong idea. Actually, we love dogs. Happiness is, of course, a warm puppy — you hardly ever see anyone fondling cold, dead puppies. Or if you do, you keep well away.

7. Suppose that the location of a mass m along an x axis is given as a function of the time t by

$$x = \ell(1 - e^{-\alpha t})$$

where ℓ and α are positive constants.

(a) How do you know that a force is acting on the mass?

(b) Determine, as a function of time, the power that this force is putting into the mass.

(c) Determine the work done on the mass by the force between times $t = 0$ and t.

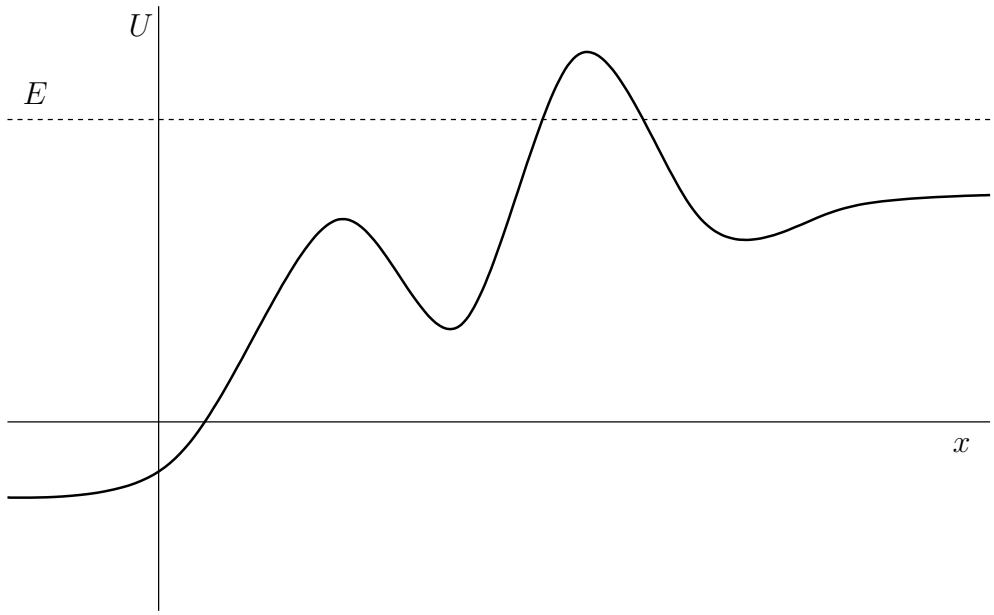


Figure 5.4: Problem 8

8. Some stupid object, starting from far down the negative x axis and moving in the positive x direction, experiences a force, the potential energy U of which is plotted in fig. (5.4) as a function of x. The dotted line on the plot corresponds to the total energy E of the object. To the left and right of the plot the potential energy U asymptotically levels off as suggested in fig. (5.4). Describe the motion of the object.

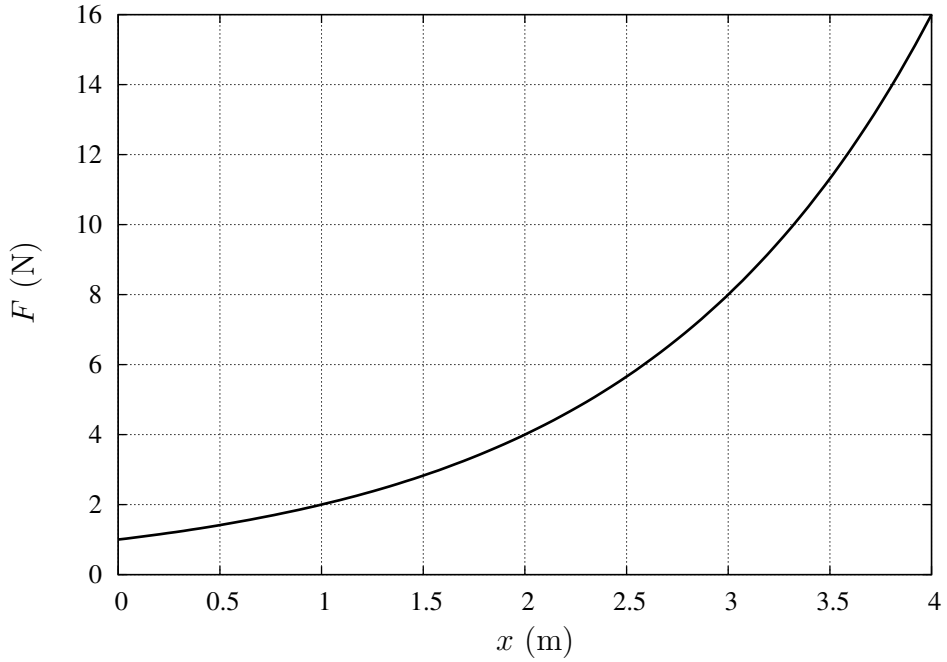


Figure 5.5: Problem 9

9. Another stupid object, starting at the origin with an initial kinetic energy of 7 J, is moving along an x axis in the positive x direction through a region where it is acted on by a retarding force (that is, a force opposite to its direction of motion). Fig. (5.5) shows a plot of the magnitude of this retarding force as a function of x. Determine approximately how far the object travels as it comes to rest. Do this graphically, without making any assumptions about the analytic expression for $F(x)$.

10. You (mass M) get hit in a very bad place by a lacrosse ball (mass m) traveling at speed V. We will suppose, very unrealistically, that in response to being hit you compress like a harmonic spring.

    (a) If the ball sinks a distance ℓ into you as it comes to rest, what is your spring constant?

    (b) As you decompress, the ball is popped back out.

        i. At what speed, ideally, would the ball pop back out?

        ii. In reality, the speed at which the ball pops back out is much less than this. Why?

11. (From *The Gods Must Be Crazy.*) While flying over the Kalahari Desert you drop an empty Coke bottle out of the window. At the instant that you drop the bottle, the plane is moving at speed V.

    (a) If the bottle descends through a height h, how fast is it moving when it beans an innocent pedestrian on the ground?

    (b) Is it also possible to arrive at this result by the methods of projectile motion?

12. You are descending a mountain pass at speed V when your brakes fail completely and, by a curious coincidence, your gearshift gets stuck in neutral. All you can do is steer and admire the scenery as you wend your way down the rolling hills.

    (a) How fast are you moving when you run a red light a distance h vertically below this point?

    (b) By some miracle, you make it through the red light, only to find that you are unable to navigate a sharp turn. You skid off the road and over a small embankment tilted up at angle θ to the horizontal, getting a spectacular view as you plummet through the air to the bottom of a gorge of depth H. At what speed are you moving when you smash into the rocks at the bottom?

    (c) Why didn't you have to take the "lay" of the road (that is, the turns and rolling hills) into account when you did these calculations?

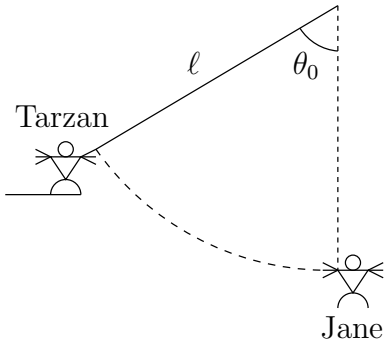


Figure 5.6: Problem 13

13. Tarzan (mass m) swings from rest on a massless vine down to Jane (also mass m), who awaits Tarzan with open arms directly under the vine's pivot point, as shown in fig. (5.6).

    (a) Determine the tension in the vine when Tarzan reaches Jane.

    (b) Why did you not have to worry about the tension in the vine when applying conservation of energy?

14. Suppose you want to determine the muzzle velocity of a spring gun. The gun is oriented horizontally, has a spring of spring constant k that is initially compressed a distance ℓ, and fires a projectile of mass m.

    Working by energy, it would seem that the initial spring potential energy $\frac{1}{2}k\ell^2$ should be converted into the projectile's kinetic energy $\frac{1}{2}mv^2$, and setting $\frac{1}{2}k\ell^2 = \frac{1}{2}mv^2$ yields

    $$v = \sqrt{\frac{k}{m}}\,\ell$$

    Working by forces, it would seem that the compressed spring exerts a force of magnitude $k\ell$ on the mass, so that $F = ma$ gives

    $$a = \frac{k}{m}\,\ell$$

    Then $v^2 - v_0^2 = 2a(x - x_0)$ becomes

    $$v^2 = 2\left(\frac{k}{m}\,\ell\right)(\ell)$$

    which yields

    $$v = \sqrt{\frac{2k}{m}}\,\ell$$

    WTF!? Which argument is wrong, and precisely how is it wrong? Or is the actual muzzle velocity a quantum superposition of these two possibilities?

15. At the company test range, an engineer repeatedly fires a standard test kitten (mass m) out of a big spring gun of mass M. The gun is in a fixed mount a height h above the conveniently level ground of the range and is aimed horizontally. It is found that when the spring gun is cocked (that is, compressed) a distance ξ, the kitten "hits the dirt" a horizontal distance ℓ from the gun, sending up cute little mushroom clouds of dust as it bounces and skips along.

    (a) From the time it leaves the gun to first impact, how long is the kitten in the air?

    (b) What was the kitten's "muzzle" velocity (that is, the speed at which the kitten left the spring gun)?

    (c) What is the spring constant of the gun?

    (d) Express the distance ℓ' that the kitten will travel when the spring gun is cocked a distance ξ' in terms of ξ' and other given quantities.

    (e) Make physical sense of this result. Once you have, you should see that it was actually possible to arrive at it without any calculation.

    (f) Would your result for #15d still hold if the gun were fired at an angle rather than horizontally?

16. Suppose that as you (mass m) go around the inside of a loop-the-loop of radius r on a roller coaster there is negligible friction, so that only gravity is changing your speed. Show that you feel a difference of $6g$'s between the bottom and the top of the loop.[14]

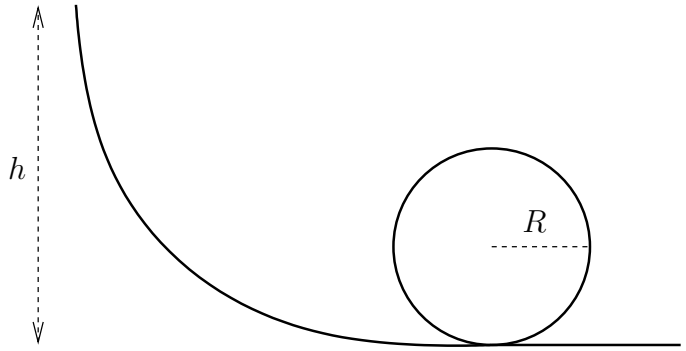


Figure 5.7: Problem 17

17. You construct a track with a loop-the-loop for your ——— trucks,[15] as shown in fig. (5.7). From what height h would you have to release the truck from rest in order for it to just barely make it around the top of the loop-the-loop? [16]

[14] And so even if you go into the loop-the-loop with just barely enough speed to make it around the top, the number of g's you experience at the bottom would still be enough to cause people to lose consciousness, have strokes, etc. While this would make for a more thrilling ride and entertaining photos, unfortunately there are other considerations, and so real loop-the-loops are not circles but clothoid loops that are sort of egg-shaped, with the small end of the egg pointing up. The effect is that the radius of curvature is smaller at the top of the loop and larger at the bottom, to compensate for your slower and faster speeds around those respective parts of the loop.

[15] Do they still make these? Ah, the memories. Of course, if they do still make them, they're probably all sissified these days: in our day, they were stamped metal, with lots of sharp edges that would get all rusty. And we were glad for those sharp, rusty edges, too. Not like kids these days.

[16] Having arrived at a result for this critical height, you would do one of three things: You would be sure to release the truck from a lower height. Or you would release the truck from a greater height, but along the track you would put GI Joes in critical positions (such as mooning the oncoming vehicle). Or you would forget about the height altogether and load the truck with firecrackers.

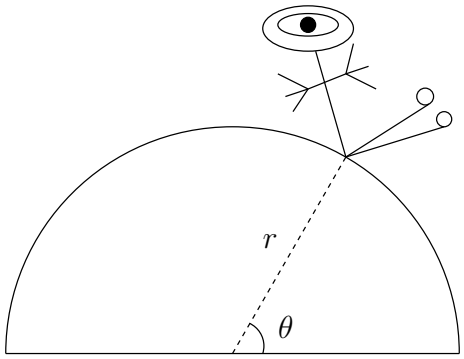


Figure 5.8: Problem 18

18. You (mass m) start sliding from rest from the top of a frictionless hemisphere of ice of radius r, as shown in fig. (5.8). (Of course, you would need a little nudge to get going, but you start essentially from rest.) At what angle θ do you lose contact with the ice and go airborne? See the footnote if you need a hint.[17]

19. Suppose that, having a spring of unknown spring constant and unknown relaxed length, and wanting to determine these two parameters in an unnecessarily complicated way, you suspend the spring vertically and hang a series of three weights from the end of it. When you hang masses m_1, m_2, and m_3 from the spring, you find that the *total* length of the spring is ℓ_1, ℓ_2, and ℓ_3, respectively.

    (a) Determine the spring constant and the relaxed length of the spring.

    (b) What relation should hold among m_1, m_2, m_3, ℓ_1, ℓ_2, and ℓ_3?

20. Fig. (5.9) shows a device that, as H.G. Wells's prediction of the division of society into *Eloi* and *Morlocks* comes ever closer to being realized, probably few of you have ever seen and even fewer ever used: a block and tackle.[18] The upper and lower rectangles (which have, so that you can see what's going on, been greatly exaggerated in size) represent the *block* parts of the *block and tackle*: in these blocks are mounted pulleys. The upper block is anchored to some structure like a ceiling by means of the green cord; from the lower block the load to be lifted (in this case, a mass m) is suspended by the blue cord. The *tackle* part of the *block and tackle* is the red cord slung around the pulleys, one end of which is fixed (tied) to the upper block. To raise the load,

[17] You will need to use both conservation-of-energy and force relations. Think about what condition on the forces corresponds to losing contact with the surface. Also, remember that you are in circular motion.

[18] Though those of you among the *Eloi* may have seen some of your servants or other proletarians using it.

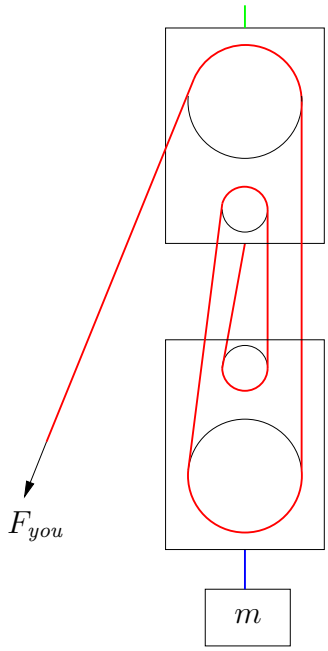


Figure 5.9: Problem 20

you pull with a force F_{you} on the loose end of the red cord. For simplicity, we will assume that the blocks and pulleys are massless and frictionless and that, in spite of the way they are drawn in fig. (5.9), all straight segments of rope are vertical.

Whew! Now that we've got all that out of the way,

(a) Reasoning by forces (as opposed to by energy), determine the force F_{you} you must apply in order to slowly raise the load — *slowly* meaning with essentially zero velocity, so that you aren't imparting any kinetic energy to the load. See the footnote if you need a hint.[19]

(b) How much work is therefore done by this force F_{you} when you lift the mass m through a height h?

(c) What is the corresponding change in the gravitational potential energy of the mass? See the footnote if you need a hint.[20]

You should have found that everything was consistent and that energy was nicely conserved. If not, you did something wrong.

[19] Note carefully how the cord is connected to the pulleys and how the tension therefore acts on the lower block in particular.

[20] Again note how the cord is wrapped around the pulleys: how far does the mass rise when you pull a length ℓ of cord at your end? This relation for distances will carry over to velocities and accelerations.

21. The Earth has a mass $m_e = 5.9723 \times 10^{24}$ kg and a radius $r_e = 6.378140 \times 10^6$ m; the Moon a mass $m_m = 7.347673 \times 10^{22}$ kg and a radius $r_m = 1.7360 \times 10^6$ m; and the center-to-center distance between the Earth and the Moon is $\ell = 3.84403 \times 10^8$ m.[21] How much work does it take to transport each kilogram of a ship (in other words, a mass $m = 1$ kg) from the surface of the Earth to the surface of the Moon?

22. Escape velocity is the critical speed with which an object would have to be launched from the surface of a planet to just barely escape the planet's gravity, in the sense that the object's speed would drop to zero as its distance from the planet became infinite.

    (a) Determine the escape velocity for the Earth, which has mass $m_e = 5.9723 \times 10^{24}$ kg and radius $r_e = 6.378140 \times 10^6$ m.

    (b) How long would it take an object projected away from the Earth with this escape velocity to make it infinitely far from the Earth?

    (c) How long would it take an object projected away from the Earth with twice this escape velocity to make it infinitely far from the Earth?

23. Finding yourself stuck on the same asteroid as the Little Prince, you immediately set about getting rid of him by heaving the little $****$ out into the void of space. On your first try, you find that when heaved at speed V radially away from the surface of the asteroid, which is a sphere of radius R, the little $****$ rises to a height h above the surface and then falls back onto the asteroid. At what minimal speed must you heave the little $****$ in order that he never return? Although it might at first seem otherwise, you do not need to know the mass of the asteroid.

[21] Actually, the orbit of the Moon has a significant eccentricity; this figure is just an average orbital radius.

24. As an alternative derivation of the result for the escape velocity of an object of mass m from a planet of mass M and radius R, someone reasons that since at the surface of the planet the gravitational force on the object is

$$F = \frac{GMm}{R^2}$$

its acceleration due to gravity at the surface is

$$g = \frac{F}{m} = \frac{GM}{R^2}$$

Using this result for the planet's g together with conservation of energy then gives the correct result for the escape velocity:

$$\tfrac{1}{2}mv_{esc}^2 = mgy = m\,\frac{GM}{R^2}\,R \qquad\qquad v_{esc} = \sqrt{\frac{2GM}{R}}$$

Find *all* of the flaws in this derivation.

25. Suppose that in some weird parallel universe, the gravitational force that a planet exerts on an object drops off exponentially with the radial distance r from the center of the planet, so that the gravitational force law is of the form

$$F_{grav} = \mathcal{G}Mm\,e^{-r/a}$$

where $\mathcal{G}$ is some weird gravitational constant, a is the radius of the planet, and M and m are the masses of the planet and the object, respectively.

(a) Determine the potential energy corresponding to this weird law of gravity. (You may take your potential energy to be zero wherever you find convenient, as long as you make clear where that is.)

(b) Determine how high an object thrown straight up with initial velocity v_0 will rise (that is, the height at the apex).

(c) Determine the escape velocity from the planet. (See #22 if you've forgotten what escape velocity is.)

(d) Without actually doing out the calculation, explain how you would determine the time it takes the object of #25b to reach the apex of its trajectory.

26. (Based on a true story.) As a kind of lame approximation of bungee jumping, some of the denizens of a dorm jump out of a third-floor window while others hold a fireman's net below. You (60 kg) drop yourself (approximately from rest) out of the window, 7.0 m above the net. We will make the crude approximation that the net acts like a harmonic spring of spring constant 3000 N/m.

    (a) How far would the "spring" (that is, the net) stretch as you are brought to rest?

    (b) You should have found that algebraically there were two solutions to the preceding part, the other being, depending on your sign conventions, ± 1.5 m. To what does this other solution correspond physically?

    (c) How far has the net stretched when you reach your maximal speed?

    (d) Unfortunately, your friends are holding the net only 1.0 m above the asphalt. How fast are you still moving when you hit the asphalt?

27. (Based on a true story.) As a kind of lame approximation of bungee jumping, some of the denizens of a dorm jump out of a third-floor window while others hold a fireman's net below. You (mass m) drop yourself (approximately from rest) out of the window, which is a height h above the net. We will make the crude approximation that the net acts like a harmonic spring of spring constant k.

    (a) How far would the "spring" (that is, the net) stretch as you are brought to rest? Be sure to make your sign conventions for directions clear.

    (b) You should have found that algebraically there were two solutions to the preceding part. To what does this other solution correspond physically?

    (c) You should also have found that these two solutions depend on m, g, and k only in the combination mg/k. Make physical sense of your solutions in the limit of

        i. Large or small h.

        ii. Large or small mg/k.

    (d) How far has the net stretched when you reach your maximal speed?

    (e) Unfortunately, the height ℓ that your friends are holding the net above the ground is less than that needed to completely stop your fall. How fast are you still moving when you hit the asphalt?

28. Suppose that when you (mass m) are butt-sledding down a hill at angle θ to the horizontal there is in fact significant friction and that after sliding a distance ℓ down it from rest your velocity is v.

    (a) How can the conservation-of-energy relation $K+U = K+U$ be modified to take into account friction? See the footnote if you need a hint.[22]

    (b) Assuming that the frictional force is constant, use your modified conservation-of-energy relation to solve for the frictional force.

29. You (mass m) are in an elevator that starts from rest and accelerates toward an upper floor with a constant acceleration a.

    (a) How can the conservation-of-energy relation $K+U = K+U$ be modified to take into account the normal force? See the footnote if you need a hint.[23]

    (b) Show explicitly that when the elevator has moved upward a distance h this modified conservation-of-energy relation yields the expected result $v = \sqrt{2ah}$ for your upward velocity.

[22] Think about how to include the work done by friction.

[23] Think about how to include the work done by the normal force.

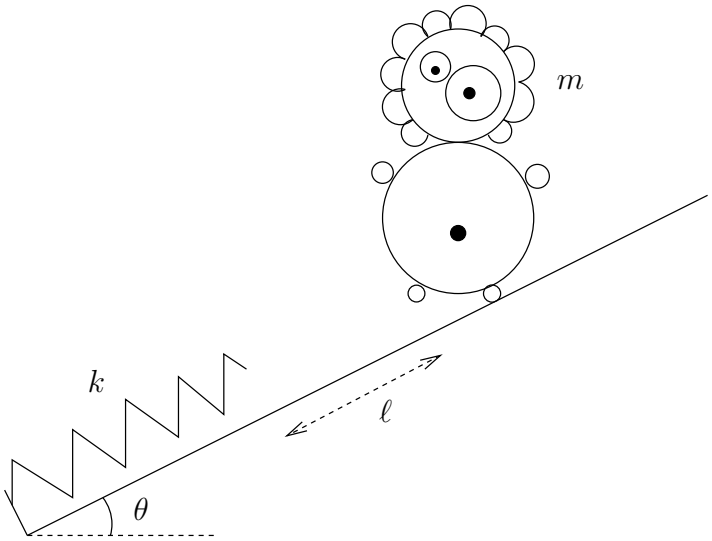


Figure 5.10: Problem 30

30. Fig. (5.10) shows a spring of spring constant k mounted to a frictionless incline at angle θ to the horizontal. To quantify how much bounce there is in the typical bouncing baby, a standard test baby of mass m is released from rest a distance ℓ from the end of the spring.

    (a) Determine how far the spring compresses as it brings the baby to rest again.

    (b) Determine the compression of the spring when the baby has reached its maximal velocity.

31. Suppose that the appeal of a bowl of potato chips falls off as $1/r^3$, where r is your distance from the bowl, so that the magnitude of the attractive force that the chips exert on you can be written

$$F = \frac{\beta}{r^3}$$

    where β is a constant.

    Determine the potential energy corresponding to this force. Make clear from where you are measuring this potential energy (that is, the point where you have put the zero of your potential).

32. Recall that the equation of motion for a mass m falling vertically under the influence of gravity and a frictional force proportional to its velocity is, if we take down to be the positive direction,

$$m\frac{dv}{dt} = mg - \beta v$$

where the frictional parameter β is a positive constant.

Suppose that after falling from rest through a distance ℓ, the mass m has essentially reached its terminal velocity. (Of course, the mass only asymptotically approaches its terminal velocity, never reaching it at any finite time, but we are supposing that after having fallen through the distance ℓ the mass is so close to its terminal velocity that it has essentially reached it.) How much work has friction then done on the mass? If you find yourself getting into a hairy calculation, you are not thinking about this the right way.

33. Starting at time $t = 0$ at $x = 0$, a mass m moves solely under the influence of a frictional force $F = -\beta v^2$, where β is a positive constant. Recall that the velocity v of the mass m is then given by

$$v = v_0 e^{-\beta x/m}$$

or, equivalently,

$$v = \frac{v_0}{1 + \beta v_0 t/m}$$

where v_0 is the velocity of the mass at $t = 0$.

(a) Using $W = \int F\,dx$, show explicitly that the work done by the frictional force in bringing the mass m to rest is $-\frac{1}{2}mv_0^2$.

(b) Using $W = \int P\,dt$, where P is the power, show explicitly that the work done by the frictional force in bringing the mass m to rest is $-\frac{1}{2}mv_0^2$.

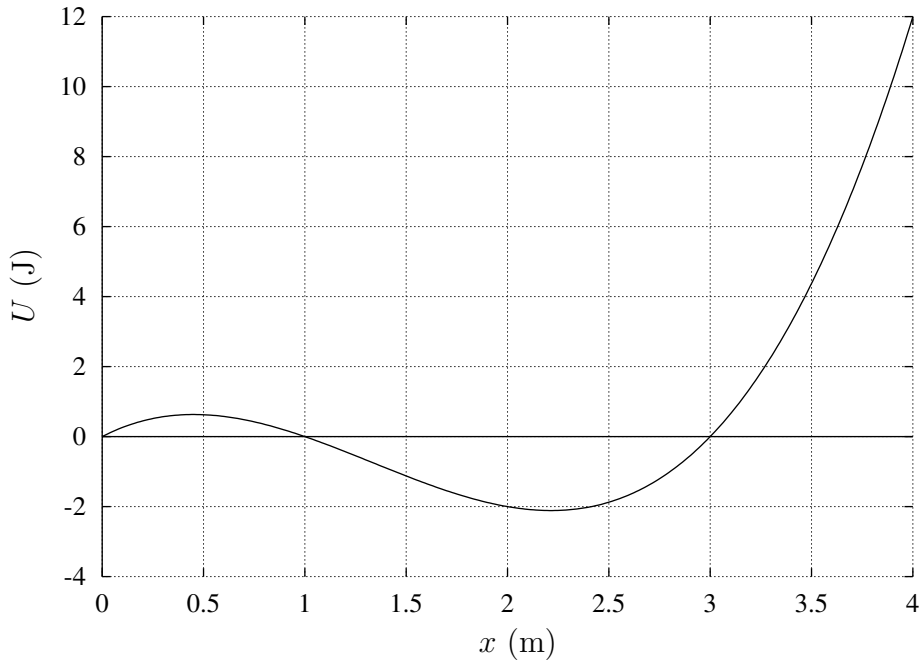


Figure 5.11: Problem 34

34. Points where the force F on a body vanishes are called equilibrium points: at these points the body is "balanced" in the sense that, there being no force on it, it will be quite happy to remain at rest there.

    (a) How can you identify such equilibrium points on a plot of a one-dimensional potential energy $U(x)$ graphically? See the footnote if you need a hint.[24]

    (b) Now that you've figured out how, identify the equilibrium points on the plot of $U(x)$ shown in fig. (5.11).

    (c) Equilibrium points are further categorized as stable or unstable. If a body is given a slight nudge away from a stable equilibrium point, the force will pull it back toward that equilibrium point; if a body is given a slight nudge away from an unstable equilibrium point, there will be a runaway effect as the force pushes it farther away.

    How can you distinguish stable and unstable equilibrium points graphically? Which of the equilibrium points in fig. (5.11) are stable and which unstable?

    (d) How therefore can you distinguish stable and unstable equilibrium points analytically (that is, mathematically)?

[24] Recall that $F = -dU/dx$.

35. A weird spring exerts a force

$$F = kx - \beta x^3$$

where k and β are positive constants and x measures the spring's distention or compression. A particle of mass m at $x = 0$ is given a very slight (that is, infinitesimal) nudge to set it in motion.

(a) Determine the potential energy corresponding to this force. You may put the zero of your potential energy wherever you want, but make clear where you are putting it.

(b) Make a rough sketch of your potential energy curve. While you do not, of course, have values for k and β, you just want to plot the general shape of the curve. And you want to do this without having to resort to your calculator.

(c) By reflecting on the force law, looking at your sketch of the potential energy, consulting your inner child, or examining the entrails of a chicken, *fully* describe the motion of the particle after it is nudged.

(d) How far out (that is, to what value of x) does the particle make it?

(e) Where does the particle reach its maximal speed?

(f) What is this maximal speed?

(g) For what range of values of x does the β term in the force law and the potential energy dominate? That is, for what range of values of x is the β term far greater in magnitude than the k term?

(h) For what range of values of x does the k term in the force law and the potential energy dominate? That is, for what range of values of x is the k term far greater in magnitude than the β term?

(i) For what range of values of x are the k and β terms in the force law and the potential energy comparable, in the sense that they are of about the same magnitude?

(j) In comparison to springs that obey the force law $F = -kx$, what is physically weird about this spring and the way it behaves?

5.7 Sketchy Answers

(7b) $-m\alpha^3\ell^2 e^{-2\alpha t}$.

(7c) $-\frac{1}{2}m\alpha^2\ell^2(1-e^{-2\alpha t})$.

(9) 2.54893723043 m.

(13a) $mg(3-2\cos\theta_0)$.

(15a) $\sqrt{\dfrac{2h}{g}}$.

(15b) $\ell\sqrt{\dfrac{g}{2h}}$.

(15c) $\dfrac{mg\ell^2}{2h\xi^2}$.

(15d) $\ell' = \dfrac{\ell}{\xi}\,\xi'$.

(17) $h = \frac{5}{2}R$.

(18) $\sin\theta = \frac{2}{3}$.

(19a) $k = \dfrac{m_i - m_j}{\ell_i - \ell_j}\,g$, relaxed length $= \ell_k - m_k\,\dfrac{\ell_i - \ell_j}{m_i - m_j}$,

where $i,\ j,\ k \in \{1,\ 2,\ 3\},\ i \neq j$.

(21) $Gm\Big[m_e\Big(\dfrac{1}{r_e} - \dfrac{1}{\ell - r_m}\Big) - m_m\Big(\dfrac{1}{r_m} - \dfrac{1}{\ell - r_e}\Big)\Big] = 5.8629 \times 10^7$ J.

(22a) $\sqrt{\dfrac{2Gm_e}{r_e}} = 11180$ m/s.

(23) $V\sqrt{\dfrac{R+h}{h}}$.

(25a) $-\mathcal{G}Mma\,e^{-r/a}$, modulo a constant, depending on where you put $U = 0$.

(25a) $-a\ln\Big(1 - \dfrac{ev_0^2}{2\mathcal{G}Ma}\Big)$.

(25a) $\sqrt{\dfrac{2\mathcal{G}Ma}{e}}$.

(26a) 1.9 m.

(26c) 0.20 m.

(26d) 10 m/s. Major bummer.

(27a) The distance would be $\dfrac{mg}{k}\Big(1 + \sqrt{1 + \dfrac{2hk}{mg}}\Big)$.

(27d) $\dfrac{mg}{k}$.

(27e) $\sqrt{2g(h+\ell) - \dfrac{k\ell^2}{m}}$.

(28b) $m\left(g\sin\theta - \dfrac{v^2}{2\ell}\right)$.

(30a) $\dfrac{mg\sin\theta}{k}\left(1 + \sqrt{1 + \dfrac{2k\ell}{mg\sin\theta}}\right)$.

(35a) With the zero at $x = 0$, $U = -\frac{1}{2}kx^2 + \frac{1}{4}\beta x^4$.

(35d) $x = \pm\sqrt{\dfrac{2k}{\beta}}$.

(35e) $x = \pm\sqrt{\dfrac{k}{\beta}}$.

(35f) $v = \dfrac{k}{\sqrt{2\beta m}}$.

Chapter 6

Center of Mass & Momentum

6.1 Center of Mass

We will take the center of mass of a system to be literally a weighted average of the parts of the system.[1] To start with a simple case, suppose we have a set of discrete of mass points m_i, $i = 1, 2, \ldots, n$, distributed along an x axis at locations x_i. The location x_{cm} of the center of mass of these mass points is defined to be

$$x_{cm} = \frac{1}{M} \sum_{i=1}^{n} m_i x_i \tag{6.1}$$

where M is the total mass:

$$M = \sum_{i=1}^{n} m_i$$

Why this is a useful definition for the center of mass will remain obscure for the time being, but at least it should agree with your intuitive sense of what constitutes a center of mass. If, for example, we have two mass points, m_1 at $x_1 = 0$ and m_2 at $x_2 = \ell$, then

$$\begin{aligned} x_{cm} &= \frac{1}{m_1 + m_2} \left(m_1 x_1 + m_2 x_2 \right) \\ &= \frac{1}{m_1 + m_2} \Big(m_1 x_1(0) + m_2 \ell \Big) \\ &= \frac{m_2}{m_1 + m_2} \, \ell \end{aligned}$$

Since

$$0 \leq \frac{m_2}{m_1 + m_2} \leq 1$$

[1] Okay, not quite literally, since in eq. (6.1) the average is weighted, not by the literal weight mg, but by the mass m.

we must have $0 \le x_{cm} \le \ell$, that is, the center of mass is always somewhere between the two masses, as you would have expected. If the two masses are equal, then we have

$$x_{cm} = \tfrac{1}{2}\ell$$

which is also just what you would have expected: in this symmetric case, the center of mass is exactly halfway between the masses. And to the extent that the masses are unequal, x_{cm} will be closer to $x = 0$ (to m_1) when $m_1 > m_2$ and closer to $x = \ell$ (to m_2) when $m_2 > m_1$. If, for example, $m_2 = 2m_1$, then $x_{cm} = \frac{2}{3}\ell$: the center of mass is twice as close to m_2 as it is to m_1.

In higher dimensions, definition (6.1) becomes

$$\mathbf{r}_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i \mathbf{r}_i \tag{6.2}$$

By dotting $\widehat{\mathbf{x}}$, $\widehat{\mathbf{y}}$, etc., into both sides of this relation, you can see that, component by component, it breaks down into a set of one-dimensional relations identical to eq. (6.1):

$$x_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i x_i\,, \qquad y_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i y_i\,, \qquad \cdots$$

Suppose, for example, we have four masses arranged at the corners of a square of side $2a$, as shown in fig. (6.1). The center of mass of these four point masses is given by

$$\begin{aligned}\mathbf{r}_{cm} &= \frac{1}{m_1+m_2+m_3+m_4}\Big(m_1(a\,\widehat{\mathbf{x}} + a\,\widehat{\mathbf{y}}) \\ &\qquad + m_2(-a\,\widehat{\mathbf{x}} + a\,\widehat{\mathbf{y}}) + m_3(-a\,\widehat{\mathbf{x}} - a\,\widehat{\mathbf{y}}) + m_4(a\,\widehat{\mathbf{x}} - a\,\widehat{\mathbf{y}})\Big) \\ &= \frac{m_1 - m_2 - m_3 + m_4}{m_1+m_2+m_3+m_4}\,a\,\widehat{\mathbf{x}} + \frac{m_1 + m_2 - m_3 - m_4}{m_1+m_2+m_3+m_4}\,a\,\widehat{\mathbf{y}}\end{aligned}$$

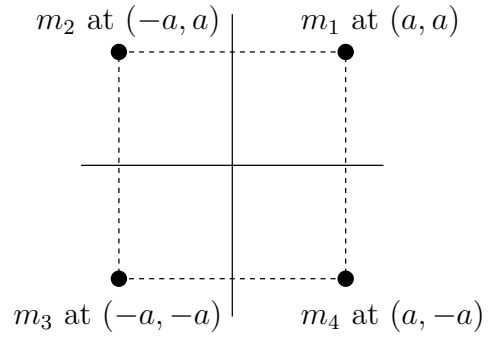


Figure 6.1: Four Masses at the Corners of a Square. *Yawn.*

In the case that the masses are all equal, this reduces to $\mathbf{r}_{cm} = 0$, as we would have expected: in this symmetric case, the center of mass is at the center of the square (the origin). In the case that $m_2 = m_4$, $\mathbf{r}_{cm}$ reduces to

$$\mathbf{r}_{cm} = \frac{m_1 - m_3}{m_1 + 2m_2 + m_3}\, a\,\hat{\mathbf{x}} + \frac{m_1 - m_3}{m_1 + 2m_2 + m_3}\, a\,\hat{\mathbf{y}}$$

which, as you can see from the equality of its x and y components, is on the line $y = x$, as we would have expected from the symmetry about that diagonal. And in the case that $m_1 = m_2 = 3m_3 = 3m_4$, $\mathbf{r}_{cm}$ reduces to

$$\mathbf{r}_{cm} = \tfrac{1}{2} a\,\hat{\mathbf{y}}$$

which, as we would have expected both from the symmetry of the masses in the $\pm x$ directions and their bias in the $\pm y$ directions, is on the y axis, three times as close to the m_1-and-m_2 side as to the m_3-and-m_4 side.

For continuous distributions of mass, the discrete sum in eq. (6.2) becomes an integral over all the infinitesimal bits of mass that make up the distribution:

$$\mathbf{r}_{cm} = \frac{1}{M} \int_{\text{distribution}} dm\, \mathbf{r} \tag{6.3}$$

where the total mass M of the distribution is just the integral over all the bits of mass that make it up:

$$M = \int_{\text{distribution}} dm \tag{6.4}$$

The expression for dm will depend on the dimension of the distribution: If the mass is distributed along a one-dimensional line, either straight or curved, then $dm = \lambda\, ds$, where λ is the linear density of the line (the mass per unit length along it) and ds is the element of arc on the line (the infinitesimal step you take as you trace it out). Similarly, if the mass is distributed along a two-dimensional surface, then $dm = \sigma\, dA$, where σ is the surface density (the mass per unit area) and dA is the area element (the area of one of the infinitesimal patches that make up the surface). And if the mass is distributed throughout a three-dimensional volume, then $dm = \rho\, dV$, where ρ is the (volume) density (the mass per unit volume) and dV is the volume element (the volume of one of the infinitesimal boxes or whatever that make up the volume). In summary:

$$dm = \begin{cases} \lambda\, ds & \text{for a linear distribution} \\ \sigma\, dA & \text{for a surface distribution} \\ \rho\, dV & \text{for a volume distribution} \end{cases}$$

The expressions for ds, dA, and dV in turn depend on the geometry of the distribution and were worked out in Chapter 2 for Cartesian, cylindrical, and spherical coordinates.

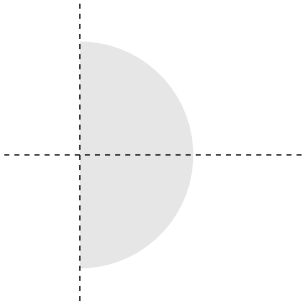


Figure 6.2: A Semicircle!

Suppose, for example, that we want to find the center of mass of the semicircular plate shown in fig. (6.2), given that the plate has radius a, mass m, and uniform density (that is, that its mass m is evenly distributed over its surface area). Since this is a two-dimensional distribution of mass, we want to use

$$\mathbf{r}_{cm} = \frac{1}{M} \int_{\text{distribution}} dm\, \mathbf{r} = \frac{1}{M} \int_{\text{plate}} \sigma\, dA\, \mathbf{r}$$

Now, M is just the mass m of the plate, and, since the plate's density is uniform, its mass per unit area σ is just $m/\frac{1}{2}\pi a^2$. The geometry of the plate being circular, it will be most natural to use polar coordinates, in which $dA = r\, dr\, d\theta$. Thus we have

$$\mathbf{r}_{cm} = \frac{1}{m} \int_{\text{plate}} \frac{m}{\frac{1}{2}\pi a^2}\, r\, dr\, d\theta\, \mathbf{r} = \frac{1}{\frac{1}{2}\pi a^2} \int_0^a r\, dr \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\theta\, \mathbf{r}$$

We now need to figure out how to express $\mathbf{r}$. We could use the polar expression $\mathbf{r} = r\,\hat{\mathbf{r}}$, which would be convenient for the integration by dr, but since the direction of $\hat{\mathbf{r}}$ varies with θ, the integration by $d\theta$ would then be relatively nasty. We will therefore use the Cartesian expression for $\mathbf{r}$:

$$\mathbf{r} = x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}}$$

This has the advantage that $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are constant and therefore independent of r and θ. But the coordinates x and y are of course not constant; to carry out the integration, we will need to express them in terms of polar coordinates by using $x = r\cos\theta$ and $y = r\sin\theta$. Thus

$$\mathbf{r} = r\cos\theta\,\hat{\mathbf{x}} + r\sin\theta\,\hat{\mathbf{y}}$$

and our integration for the center of mass becomes

$$\mathbf{r}_{cm} = \frac{1}{\frac{1}{2}\pi a^2} \int_0^a r\, dr \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} d\theta\, (r\cos\theta\,\hat{\mathbf{x}} + r\sin\theta\,\hat{\mathbf{y}})$$

$$
\begin{aligned}
&= \frac{1}{\frac{1}{2}\pi a^2}\left[\left(\int_0^a r^2\,dr\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\cos\theta\,d\theta\right)\widehat{\mathbf{x}}\right.\\
&\qquad\qquad\left.+\left(\int_0^a r^2\,dr\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\sin\theta\,d\theta\right)\widehat{\mathbf{y}}\right]\\
&= \frac{1}{\frac{1}{2}\pi a^2}\left[\left(\left[\tfrac{1}{3}r^3\right]_0^a\Big[\sin\theta\Big]_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\right)\widehat{\mathbf{x}}\right.\\
&\qquad\qquad\left.+\left(\left[\tfrac{1}{3}r^3\right]_0^a\Big[-\cos\theta\Big]_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\right)\widehat{\mathbf{y}}\right]\\
&= \frac{1}{\frac{1}{2}\pi a^2}\left[\tfrac{1}{3}a^3\,(2)\,\widehat{\mathbf{x}}+\tfrac{1}{3}a^3\,(0)\,\widehat{\mathbf{y}}\right]\\
&= \frac{4}{3\pi}\,a\,\widehat{\mathbf{x}}
\end{aligned}
$$

As we would have expected from the $\pm y$ symmetry, the center of mass is on the x axis. Also note that $\frac{4}{3\pi} < \frac{1}{2}$: since, in terms of its x coordinate, more of the mass of the plate is toward its base than out toward $x = a$, the center of mass is closer to the base than to $x = a$.

In the preceding example, we could have taken advantage of the symmetry of the distribution from the start by noting that the center of mass had to end up being on the x axis when we set up our integrations. We will make such use of symmetry in this next example: suppose we have a hemisphere of radius a, mass m, and uniform density, oriented so that its base is centered in the xy plane and it bulges up in the positive z direction, as shown in fig. (6.3). From symmetry, we know ahead of time that the center of mass has to end up being on the z axis. We need therefore concern ourselves only with the z component of the center of mass; we know that x_{cm} and y_{cm} will both vanish. We have only to calculate

$$z_{cm} = \frac{1}{M}\int\limits_{\text{hemisphere}} dm\,z$$

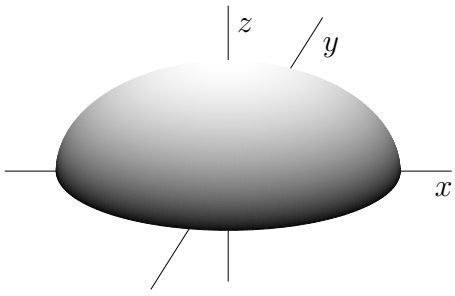


Figure 6.3: A Hemisphere. Whoopee.

M will just be the mass m of the hemisphere, and for a volume distribution the mass element dm will be $\rho\, dV$, with the (constant) density ρ being simply

$$\rho = \frac{m}{\frac{1}{2}\frac{4}{3}\pi a^3}$$

and the volume element dV being, in the spherical coordinates natural to the shape we are integrating over,

$$dV = r^2 \sin\theta\, dr\, d\theta\, d\varphi$$

To carry out the integrations, we need to express z in spherical coordinates: $z = r\cos\theta$. Putting this all together, we have

$$\begin{aligned}
z_{cm} &= \frac{1}{m} \int_{\text{hemisphere}} \frac{m}{\frac{1}{2}\frac{4}{3}\pi a^3}\, r^2 \sin\theta\, dr\, d\theta\, d\varphi\, r\cos\theta \\
&= \frac{1}{\frac{2}{3}\pi a^3} \int_0^a r^3\, dr \int_0^{\frac{\pi}{2}} \sin\theta\cos\theta\, d\theta \int_0^{2\pi} d\varphi \\
&= \frac{1}{\frac{2}{3}\pi a^3} \left(\tfrac{1}{4}a^4\right)\left(\int_0^{\frac{\pi}{2}} \tfrac{1}{2}\sin(2\theta)\, d\theta\right) (2\pi) \\
&= \frac{1}{\frac{2}{3}\pi a^3} \left(\tfrac{1}{4}a^4\right)\Big[-\tfrac{1}{4}\cos(2\theta)\Big]_0^{\frac{\pi}{2}} (2\pi) \\
&= \frac{1}{\frac{2}{3}\pi a^3} \left(\tfrac{1}{4}a^4\right)\left(\tfrac{1}{2}\right)(2\pi) \\
&= \tfrac{1}{6}a
\end{aligned}$$

Since the majority of the mass of the hemisphere is even closer to its base than was that of the semicircle, the hemisphere's center of mass is even closer to its base than the semicircle's.

Finally, we note that the center of mass can be determined for parts of a system and then combined, as when determining the center of mass of an L shape from the centers of mass of the two segments of which it is composed. Mathematically, this amounts simply to separating the integration for the center of mass into pieces: in the case of the L shape,

$$\mathbf{r}_{cm} = \frac{1}{M}\int_{\mathsf{L}} dm\, \mathbf{r} = \frac{1}{M}\left(\int_{\text{base of L}} dm\, \mathbf{r} + \int_{\text{side of L}} dm\, \mathbf{r} \right)$$

Distributions of mass with missing pieces or holes can likewise be regarded as complete distributions plus a piece with an effectively negative mass:

$$\mathbf{r}_{cm} = \frac{1}{M} \int_{\text{dist. with hole}} dm\, \mathbf{r}$$

$$= \frac{1}{M}\left[\int\limits_{\text{dist. with hole}} dm\,\mathbf{r} + \left(\int\limits_{\text{missing piece}} dm\,\mathbf{r} \;-\; \int\limits_{\text{missing piece}} dm\,\mathbf{r}\right)\right]$$

$$= \frac{1}{M}\left[\left(\int\limits_{\text{dist. with hole}} dm\,\mathbf{r} \;+\; \int\limits_{\text{missing piece}} dm\,\mathbf{r}\right) - \int\limits_{\text{missing piece}} dm\,\mathbf{r}\right]$$

$$= \frac{1}{M}\left[\int\limits_{\text{complete dist.}} dm\,\mathbf{r} \;-\; \int\limits_{\text{missing piece}} dm\,\mathbf{r}\right]$$

Consider, for example, the shape shown in fig. (6.4), which is composed of three squares, each of side a, and which we will take to conveniently have mass $3m$ evenly distributed over its area. There are two ways to find the center of mass of this plate. First, we could note that the plate is composed of three squares, each of mass m, with centers of mass located at $(-\frac{1}{2}a, \frac{1}{2}a)$, $(-\frac{1}{2}a, -\frac{1}{2}a)$, and $(\frac{1}{2}a, -\frac{1}{2}a)$. The center-of-mass calculation for the plate is therefore equivalent to that of three point masses at those locations:

$$\begin{aligned}\mathbf{r}_{cm} &= \frac{1}{3m}\Big(m\big(-\tfrac{1}{2}a\,\widehat{\mathbf{x}} + \tfrac{1}{2}a\,\widehat{\mathbf{y}}\big) + m\big(-\tfrac{1}{2}a\,\widehat{\mathbf{x}} - \tfrac{1}{2}a\,\widehat{\mathbf{y}}\big) + m\big(\tfrac{1}{2}a\,\widehat{\mathbf{x}} - \tfrac{1}{2}a\,\widehat{\mathbf{y}}\big)\Big) \\ &= -\tfrac{1}{6}\,a(\widehat{\mathbf{x}} + \widehat{\mathbf{y}})\end{aligned}$$

Alternatively, we could regard the plate as a complete square of mass $4m$, less a square of mass m in the upper right. The center of mass of the complete square would be, by symmetry, at the origin; that of the missing square on upper right would be at $(\frac{1}{2}a, \frac{1}{2}a)$. The center of mass of the plate would therefore be given by

$$\mathbf{r}_{cm} = \frac{1}{4m - m}\Big(4m(0) - m\big(\tfrac{1}{2}a\,\widehat{\mathbf{x}} + \tfrac{1}{2}a\,\widehat{\mathbf{y}}\big)\Big) = -\tfrac{1}{6}\,a(\widehat{\mathbf{x}} + \widehat{\mathbf{y}})$$

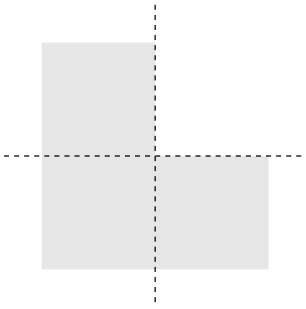


Figure 6.4: “Upon my word!” said Farmer Brown

6.2 The Dynamics of the Center of Mass

We can get results for the velocity and acceleration of the center of mass of a system of point masses simply by taking time derivatives of the position vector of the center of mass: [2]

$$\begin{aligned}\mathbf{v}_{cm} &= \frac{d\mathbf{r}_{cm}}{dt} \\ &= \frac{d}{dt}\left(\frac{1}{M}\sum_{i=1}^{n} m_i \mathbf{r}_i\right) \\ &= \frac{1}{M}\sum_{i=1}^{n} m_i \frac{d\mathbf{r}_i}{dt} \\ &= \frac{1}{M}\sum_{i=1}^{n} m_i \mathbf{v}_i \end{aligned} \tag{6.5}$$

and likewise

$$\mathbf{a}_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i \mathbf{a}_i \tag{6.6}$$

Newton's second law, $\mathbf{F} = m\mathbf{a}$, is strictly posited only for mass points. But for systems of mass points, the net force on the system would just be the sum of the net forces on each of the masses m_i that make up the system, and the net forces on the m_i do obey Newton's second law:

$$\mathbf{F}_{net,\ system} = \sum_{i=1}^{n} \mathbf{F}_{net,i} = \sum_{i=1}^{n} m_i \mathbf{a}_i$$

We can gerrymander [3] this by multiplying and dividing by the total mass M of the system and using eq. (6.6):

$$\mathbf{F}_{net,\ system} = M\left(\frac{1}{M}\sum_{i=1}^{n} m_i \mathbf{a}_i\right) = M\mathbf{a}_{cm}$$

This is telling us that $\mathbf{F} = m\mathbf{a}$ does in fact apply to systems, as long as we are careful to note that the mass involved is the total mass of the system and the acceleration is specifically that of the center of mass of the system. We can, however, simplify this still further: the net force on the system can be regarded as the sum of two types of forces: external forces exerted on the system by sources outside the system, and internal forces exerted by the various bodies in the system on each other. If we regard you as a system, for

[2] We work out our derivations only for the case of a distribution of discrete point masses, but the proofs are the same for continuous distributions and for mixed discrete and continuous distributions.

[3] *Gerrymander* is such a great word.

example, the external force of gravity that is exerted by the Earth on each of the molecules that make up your person holds you on the ground, and internal forces that the molecules of your person exert on each other hold you in your familiar shape (as opposed to collapsing into a disgusting puddle of goo on the ground). According to Newton's third law, the internal forces must come in equal and opposite pairs that will therefore cancel out when we sum to get the net force on the system. So only external forces make a nonzero contribution the net force, and we therefore have

$$\mathbf{F}_{net,\ ext} = M\mathbf{a}_{cm} \tag{6.7}$$

When applying $\mathbf{F} = m\mathbf{a}$ to systems, we therefore need consider only the external forces acting on the system. Eq. (6.7) is a hugely important result, not only in principle, but also in practice, because internal forces like those that hold rigid bodies in their shapes are generally very nasty and complicated.

One consequence of all this is that when applying the form mgy of the gravitational potential to an extended body (that is, to a body that is not simply a mass point), for the y coordinate you should use the y of the body's center of mass, since it is, according to eq. (6.7), the motion of the body's center of mass that the force mg determines.[4] We have been getting this right up to now without even thinking about it because the extended bodies with which we have dealt have not changed their orientation: when, for example, a baby was heaved into the air, we should properly have used the change in height of the baby's center of mass for Δy, but since the baby was assumed to remain right-side up (or upside down, or however it was initially oriented) throughout its flight, the Δy of the center of mass is the same as the Δy of the head or the Δy of the feet or the Δy of the bellybutton — as long as we were consistent about the part of the baby that we used to determine Δy, we were okay. But when we get to rotational dynamics and we start dealing with objects that are changing their orientation (such as babies tumbling as they fly through the air), it will be very important to use the y of the center of mass in mgy.

6.3 Momentum & Momentum Conservation

In the previous chapter, we noted that conservation of energy was a direct consequence of the invariance of physics under time translations, that because the laws of physics are the same today as they were yesterday and will be tomorrow, there necessarily must be a conserved energy. As we will see

[4] At least in part; there may of course also be other external forces acting.

A rigorous proof of all this will have to wait until §7.9, when we get to rotational dynamics.

when we do relativity, our three spatial and one time dimensions are not, as everyday experience would lead you to believe, separate from each other; they form a four-dimensional spacetime and can transform into each other: what is purely a spatial distance to one person may be partly a time interval to another, and likewise what is purely a time interval to one person may be partly a spatial distance to another. Anything that applies to time must therefore also apply to space and vice versa. So if the universe is symmetric under time translations, it must also be symmetric under space translations, that is, the laws of physics must be the same everywhere in the universe. Energy was the conserved quantity corresponding to invariance under time translations; the spatial analog of energy, the quantity that turns out to be conserved as a result of invariance under space translations, is momentum.

As we did in the case of energy, we will, however, approach things from the more pedestrian level of Newton's laws. If we were so inclined, we could rewrite $\mathbf{F} = m\mathbf{a}$ as follows:

$$\mathbf{F} = m\mathbf{a} = m\frac{d\mathbf{v}}{dt} = \frac{d(m\mathbf{v})}{dt}$$

In other words, the net force acting on a body can be regarded as the rate of change of the quantity $m\mathbf{v}$, which we define to be the momentum $\mathbf{p}$:

$$\mathbf{p} = m\mathbf{v} \tag{6.8a}$$

$$\mathbf{F} = \frac{d\mathbf{p}}{dt} \tag{6.8b}$$

This definition of momentum may not coincide entirely with your intuitive notions of momentum. In a loose, subjective way, however, $p = mv$, being proportional to both the mass and the velocity of the body, does give a measure of the body's oomph,[5] that is, of how hard you would feel you'd been hit if the body struck you.

In the case that there is no net force acting on the body, eq. (6.8b) leads to a conservation law: since its rate of change is zero, the momentum $\mathbf{p}$ of the body is constant.

Now, for a single body, this isn't telling us anything we didn't already know from $\mathbf{F} = m\mathbf{a}$ (in fact, from Newton's first law): when there is no net force on a body, its velocity is constant. And if $\mathbf{v}$ is constant, so of course is $\mathbf{p} = m\mathbf{v}$. Duh.

Conservation of momentum can, however, also be applied to systems of multiple bodies: as in the previous section, to get the net force on a system, we simply add the net forces on the bodies that make up that system, but this time we write things in terms of momentum:

$$\mathbf{F}_{net,\ system} = \sum_{i=1}^{n} \mathbf{F}_{net,i}$$

[5] Okay, so *oomph* isn't really a scientific technical term. But it should be.

$$
\begin{aligned}
&= \sum_{i=1}^{n} \frac{d\mathbf{p}_i}{dt} \\
&= \frac{d\left(\sum_{i=1}^{n} \mathbf{p}_i\right)}{dt} \\
&= \frac{d\mathbf{p}_{\textit{total, system}}}{dt}
\end{aligned}
$$

That is, the net force on the system of bodies is equal to the rate of change of the system's total momentum. Since, according to eq. (6.7), only external forces make a nonzero contribution the net force on the system, we have

$$\mathbf{F}_{\textit{net ext, system}} = \frac{d\mathbf{p}_{\textit{total, system}}}{dt} \tag{6.9}$$

This is another hugely important result: if no net external force acts on a system, then the total momentum of that system is conserved. Such systems are sometimes called isolated. The most prominent example of an isolated system is the universe itself: there is nothing outside of it to exert any forces on it. Just as conservation of energy meant that energy could only be redistributed among its various forms, conservation of momentum means that momentum can only be redistributed among the bodies in the system; the total remains constant. So the total momentum of the universe, like its total energy, is conserved.[6]

For the momentum of a system to be exactly conserved, there must be no net external force on it. But there is a commonly occurring case for which the momentum of a system is conserved to a more or less good approximation even when there is a net external force: if we rewrite eq. (6.9) as

$$d\mathbf{p}_{\textit{total, system}} = \mathbf{F}_{\textit{net ext, system}}\, dt$$

and integrate

$$
\begin{aligned}
\int d\mathbf{p}_{\textit{total, system}} &= \int \mathbf{F}_{\textit{net ext, system}}\, dt \\
\Delta\mathbf{p}_{\textit{total, system}} &= \int \mathbf{F}_{\textit{net ext, system}}\, dt
\end{aligned}
$$

[6] Actually, we should note, for the benefit of those of you whose psychology is such that you are not happy unless you have some vaguely disturbing doubt gnawing at the pit of your soul, that in general relativity it remains to this day difficult even to determine a physically and mathematically viable definition of energy and momentum. A single, comprehensive construction that applies more generally than to certain geometries or for isolated systems, and in particular a local conservation law, remains elusive. See, for example, Bamba & Shimizu, *Construction of energy-momentum tensor of gravitation*, `https://arxiv.org/abs/1506.02760`, and references therein.

we see that the change in the total momentum of the system will be small, and consequently momentum will be approximately conserved, as long as the time interval over which we are integrating is too short for the net external force to cause a significant change in momentum. Most collisions occur very quickly, so that even if there is a substantial net external force on the system, the total momentum immediately after the collision does not differ significantly from the total momentum immediately before the collision. When, for example, a missile explodes in midair, that physical system (the missile) is experiencing the very substantial external forces of gravity and air resistance, but the explosion occurs so quickly that the total momentum of the missile just before the explosion can to a good approximation be equated to the total momentum of its various pieces just after the explosion.

6.4 Collisions

In practice, you will usually apply conservation of momentum by equating the total momentum of a system before and after some process, usually a collision or explosion. Suppose, for example, you try to be funny by riding your skateboard straight at your best friend, but your friend gets the last laugh by pulling out a sawed-off shotgun loaded with deer slugs and blowing you away. Your getting shot constitutes a collision between the parts of a two-body system consisting of you (including the skateboard) and the slug from the shotgun. The total momentum of the system is simply the sum of your momentum and the slug's, and you therefore conserve momentum by equating the value of this total just before you get hit to its value just after you get hit:

$$m_{you}v_{you,\ before} + m_{slug}v_{slug,\ before} = m_{you}v_{you,\ after} + m_{slug}v_{slug,\ after}$$

If the slug sticks inside you, then afterward you and the slug share the same velocity, and this reduces to

$$m_{you}v_{you,\ before} + m_{slug}v_{slug,\ before} = m_{total}v_{both,\ after}$$

Even though this is just a one-dimensional example, don't forget that the v's are of course velocities: momentum is a vector quantity, and even in one dimension we need to include the $\pm$ signs for the directions. If you were originally moving in the positive direction, then the bullet would have a negative velocity as it approached you.

All collisions for which momentum is conserved (either exactly because there is no net force or approximately because the collision occurs quickly) are further classified by what happens to the total kinetic energy of the system. The system's kinetic energy can remain constant or, more usually,

either decrease or increase. By definition, if the kinetic energy changes, the collision is inelastic; if the kinetic energy remains the same, it is elastic.[7] Most commonly, at least some kinetic energy is lost: it takes energy to shatter glass, twist metal, break bones, etc., and some energy is almost always lost to internal friction as objects deform. The macroscopic kinetic energy lost to internal friction reappears in the microscopic form of random molecular motion, that is, heat. In §6.5 we will see that the worst case is when bodies stick together: the loss of kinetic energy is then maximal, and such collisions are therefore termed completely inelastic.[8] Kinetic energy can also increase in a collision — as, for example, when it is released in an explosion. Barring such a release of energy, in order for a collision to be nearly or exactly elastic there must be little or no loss of kinetic energy to damage or internal friction during deformations. This is the case with hard, rigid bodies, like billiard balls, or with special polymers like those used to make "bouncy balls."

If you know that a collision is elastic, you can and should equate the total kinetic energy before and after the collision. For a one-dimensional collision between two bodies, you would therefore have

$$\tfrac{1}{2}mv_{1,\,before}^2 + \tfrac{1}{2}mv_{2,\,before}^2 = \tfrac{1}{2}mv_{1,\,after}^2 + \tfrac{1}{2}mv_{2,\,after}^2$$

Do not, however, assume that a collision is elastic unless you have some definite reason to believe that it is; bad assumptions of this sort are a very common error.

6.5 Center-of-Mass & Relative Coordinates

We will denote by $\mathbf{r}_i$ the position vector of the point mass m_i. This position vector, which by definition goes from the origin to the location of m_i, can be rewritten as the sum

$$\mathbf{r}_i = \mathbf{r}_{cm} + \mathbf{r}_{i,cm} \tag{6.10}$$

where $\mathbf{r}_{cm}$, which goes from the origin to the center of mass, is the position vector of the center of mass, and where $\mathbf{r}_{i,cm}$, which goes from the center of mass to m_i, is the position vector of m_i *relative to the center of mass*, as shown in fig. (6.5).

[7] The *total* energy is, of course, always conserved; it is just that there is a redistribution of the forms in which this total energy appears.

[8] Actually, life will get more complicated later on when we deal with rotational motion: it will turn out that, as you might expect, the rotational motion of a system makes a contribution to its kinetic energy, and even when objects stick together they may, depending on just how they struck each other, be rotating after the collision. But for the present we will not worry about such rotational effects and life will be blissfully simple.

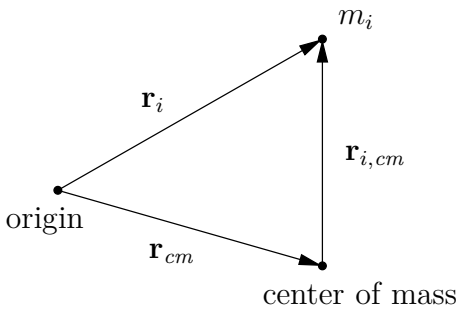


Figure 6.5: A Recipe for Confusion

Taking the derivative of eq. (6.10) with respect to time gives us the corresponding relation among velocities:

$$\mathbf{v}_i = \mathbf{v}_{cm} + \mathbf{v}_{i,cm} \tag{6.11}$$

where $\mathbf{v}_i$ is the velocity of m_i, $\mathbf{v}_{cm}$ is the velocity of the center of mass, and $\mathbf{v}_{i,cm}$ is the velocity of m_i relative to the center of mass.

Recall now eq. (6.5),

$$\mathbf{v}_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i \mathbf{v}_i$$

which tells us that the velocity of the center of mass is just the same weighted average of velocities of the m_i that the location of the center of mass is of the locations of the m_i. Expressing this is terms of momentum, we have

$$\mathbf{v}_{cm} = \frac{1}{M}\sum_{i=1}^{n} \mathbf{p}_i = \frac{1}{M}\mathbf{p}_{\textit{total, system}}$$

or, in other words,

$$\mathbf{p}_{\textit{total, system}} = M\mathbf{v}_{cm} \tag{6.12}$$

That is, the total momentum of the system can equivalently be regarded as the momentum you would have if the whole system were collapsed down to a point mass M at the center of mass. This is yet another hugely important result.

A reference frame is just the perspective of an observer. If you are driving a car down the positive x axis, you and the car and everything in the car share the same reference frame — which differs, for example, from the reference frame of people standing at rest on the ground. From their perspective, you are the one who is moving, while from your own perspective you and the car are at rest and it is they and the trees and phone poles that are moving in the negative x direction.

The center-of-mass frame is defined to be the reference frame moving with the center of mass. That is, the reference frame of the center of mass is just

the perspective of an observer who shares the same velocity as the center of mass. In particular, if you are in the center-of-mass frame, you are moving with the center of mass, so that from your perspective $\mathbf{v}_{cm} = 0$. According to eq. (6.12), this means that from your perspective $\mathbf{p}_{\textit{total, system}} = 0$ as well. The center-of-mass frame could therefore equally well be called the center-of-momentum frame.

The total kinetic energy in *any* reference frame will be given by

$$K_{total} = \sum_{i=1}^{n} K_i = \sum_{i=1}^{n} \tfrac{1}{2} m_i v_i^2$$

which, if we make use of eq. (6.11) and also the general vector relation $a^2 = \mathbf{a} \cdot \mathbf{a} = \mathbf{a}^2$, becomes

$$\begin{aligned}
K_{total} &= \sum_{i=1}^{n} \tfrac{1}{2} m_i (\mathbf{v}_{cm} + \mathbf{v}_{i,cm})^2 \\
&= \sum_{i=1}^{n} \tfrac{1}{2} m_i (v_{cm}^2 + 2\mathbf{v}_{cm} \cdot \mathbf{v}_{i,cm} + v_{i,cm}^2) \\
&= \sum_{i=1}^{n} \tfrac{1}{2} m_i v_{cm}^2 + \sum_{i=1}^{n} m_i \mathbf{v}_{cm} \cdot \mathbf{v}_{i,cm} + \sum_{i=1}^{n} \tfrac{1}{2} m_i v_{i,cm}^2 \\
&= \tfrac{1}{2} \left(\sum_{i=1}^{n} m_i \right) v_{cm}^2 + \mathbf{v}_{cm} \cdot \left(\sum_{i=1}^{n} m_i \mathbf{v}_{i,cm} \right) + \sum_{i=1}^{n} \tfrac{1}{2} m_i v_{i,cm}^2 \\
&= \tfrac{1}{2} \left(\sum_{i=1}^{n} m_i \right) v_{cm}^2 + \mathbf{v}_{cm} \cdot M \left(\frac{1}{M} \sum_{i=1}^{n} m_i \mathbf{v}_{i,cm} \right) + \sum_{i=1}^{n} \tfrac{1}{2} m_i v_{i,cm}^2
\end{aligned}$$

According to eq. (6.5), the parenthetic expression in the middle term is just the velocity of the center of mass *relative to the center of mass*, which will vanish: the center of mass isn't moving relative to itself. This kills the contribution of the middle term. The parenthetic expression in the first term is just the total mass M of the system. We thus arrive at

$$K_{total} = K_{cm} + K_{rel} \tag{6.13a}$$

where we have defined

$$K_{cm} = \tfrac{1}{2} M v_{cm}^2 \tag{6.13b}$$

and

$$K_{rel} = \sum_{i=1}^{n} \tfrac{1}{2} m_i v_{i,cm}^2 \tag{6.13c}$$

That is, the total kinetic energy of a system can always be regarded as the sum of two contributions: K_{cm}, which is the kinetic energy *of* the center of mass (in the sense that it is the kinetic energy you would have if the system

were collapsed down to a mass point at the center of mass), and K_{rel}, which is the kinetic energy *relative to* the center of mass (in the sense that it is the sum of kinetic energies due to the velocities $\mathbf{v}_{i,cm}$ of the m_i relative to the center of mass). Eqq. (6.13a)-(6.13c) are hugely important. And it will turn out that this very neat and useful decomposition into center-of-mass and relative contributions occurs with other physical quantities as well.

In the context of collisions, this means that whenever momentum is conserved, K_{cm} will remain constant: constant total momentum means, according to eq. (6.12), constant $\mathbf{v}_{cm}$, which in turn, according to eq. (6.13b), means constant K_{cm}. Since K_{cm} is constant in any collision that conserves momentum, the only contribution to the total kinetic energy that can change is, according to eq. (6.13a), the relative kinetic energy K_{rel}. And since the minimal value of any kinetic energy is zero, this means that the loss of kinetic energy in a collision is greatest when all of the initial K_{rel} is lost. From eq. (6.13c), we can see that losing all of the initial relative kinetic energy (having a final $K_{rel} = 0$) means that every contribution $\frac{1}{2}m_i v_{i,cm}^2$ to the final K_{rel} must be zero, that is, that no m_i has any velocity relative to the center of mass. In other words, the loss of kinetic energy is greatest when all the parts of the system stick together, thereby losing all of their relative motion. This is why collisions in which bodies stick together are termed completely inelastic. Physically, it is momentum conservation that gives rise to this upper limit on the loss of kinetic energy: if everything were to come to rest, so that *all* of the system's kinetic energy, both center-of-mass and relative, were lost, the total momentum after the collision would vanish, and this would in general violate conservation of momentum.

In addition to the special case of complete inelasticity, you should remember the more general results that in collisions that conserve momentum

- the velocity $\mathbf{v}_{cm}$ of the center of mass does not change, and
- the only part of the system's kinetic energy that can change, whether that change is an increase or decrease or zero, is the relative part, K_{rel}.

In the center-of-mass frame, because $\mathbf{v}_{cm} = 0$, $K_{cm} = 0$: in the center-of-mass frame, the only contribution to the kinetic energy is the relative contribution.

6.6 Rockets

Rockets are an example of a change in momentum $\mathbf{p} = m\mathbf{v}$ due to a change, not in the velocity $\mathbf{v}$, but in the mass m. If, for example, a rocket in empty space hurls mass (combusted fuel) backward at a constant speed u relative to the rocket itself, conservation of the total momentum of the rocket-fuel system requires that the gain in forward momentum of the rocket equals the

backward momentum of the ejected fuel. For a rocket of mass m experiencing an increase dv in its velocity v, the boost in the rocket's forward momentum is $m\,dv$. And since the ejected fuel is moving backward at speed u, its contribution to the change in the momentum of the rocket-fuel system will be $-u\,dm_{fuel}$. Setting the total change in momentum to zero, we have

$$m\,dv - u\,dm_{fuel} = 0$$

Now, the mass dm_{fuel} of the ejected fuel literally comes out of the rocket's mass, so that $dm_{fuel} = -dm$, where the negative sign is to take into account that the rocket is losing mass ($dm < 0$). Our momentum-conservation relation thus becomes

$$m\,dv + u\,dm = 0$$

Separating variables and integrating then yields

$$dv = -u\,\frac{dm}{m}$$

$$\int_{v_0}^{v} dv = -u\int_{m_0}^{m}\frac{dm}{m}$$

$$v - v_0 = -u\ln\frac{m}{m_0} = u\ln\frac{m_0}{m}$$

So much for rockets.

6.7 Two-Body Collisions

Lab Coordinates

The term lab frame is used in two senses. Usually it refers to the reference frame in which one of the colliding bodies is at rest, since in an actual lab one commonly fires objects at a stationary target. More generally, any frame other than the center-of-mass frame can be called a lab frame.

In this section we will restrict ourselves to collisions between two bodies, and we will use numerical subscripts to distinguish the various physical quantities associated with the two masses and primes to denote the value of quantities after the collision. Thus our notation for momentum conservation will be

$$m_1\mathbf{v}_1 + m_2\mathbf{v}_2 = m_1\mathbf{v}'_1 + m_2\mathbf{v}'_2 \tag{6.14}$$

and, when the collision is elastic, for conservation of kinetic energy it will be

$$\tfrac{1}{2}m_1v_1^2 + \tfrac{1}{2}m_2v_2^2 = \tfrac{1}{2}m_1v'^2_1 + \tfrac{1}{2}m_2v'^2_2 \tag{6.15}$$

You should be able to set up and solve collision problems in this conventional frame. You should also be able to set up and solve collision problems in the center-of-mass frame:

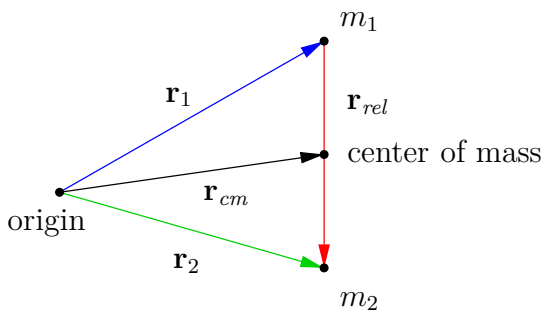


Figure 6.6: Relative & Center-of-Mass Coordinates

Center-of-Mass & Relative Coordinates

When we have just two bodies, instead of the locations $\mathbf{r}_1$ and $\mathbf{r}_2$ of each body, we can work with the location $\mathbf{r}_{cm}$ of the center of mass and $\mathbf{r}_{rel}$, the location of m_2 relative to m_1, as shown in fig. (6.6): [9]

$$\begin{aligned} \mathbf{r}_{rel} &= \mathbf{r}_2 - \mathbf{r}_1 \\ \mathbf{r}_{cm} &= \frac{1}{m_1 + m_2}(m_1\mathbf{r}_1 + m_2\mathbf{r}_2) \end{aligned} \tag{6.16}$$

Algebraically inverting these relations yields

$$\begin{aligned} \mathbf{r}_1 &= \mathbf{r}_{cm} - \frac{m_2}{m_1 + m_2}\mathbf{r}_{rel} \\ \mathbf{r}_2 &= \mathbf{r}_{cm} + \frac{m_1}{m_1 + m_2}\mathbf{r}_{rel} \end{aligned} \tag{6.17}$$

If we use $\mathbf{r}_i = \mathbf{r}_{cm} + \mathbf{r}_{i,cm}$ from eq. (6.10), the locations $\mathbf{r}_{1,cm}$ and $\mathbf{r}_{2,cm}$ of m_1 and m_2 relative to the center of mass, shown in fig. (6.7), are therefore [10]

$$\begin{aligned} \mathbf{r}_{1,cm} &= \mathbf{r}_1 - \mathbf{r}_{cm} \\ &= \left(\mathbf{r}_{cm} - \frac{m_2}{m_1 + m_2}\mathbf{r}_{rel}\right) - \mathbf{r}_{cm} \\ &= -\frac{m_2}{m_1 + m_2}\mathbf{r}_{rel} \\ \mathbf{r}_{2,cm} &= \mathbf{r}_2 - \mathbf{r}_{cm} \\ &= \left(\mathbf{r}_{cm} + \frac{m_1}{m_1 + m_2}\mathbf{r}_{rel}\right) - \mathbf{r}_{cm} \\ &= \frac{m_1}{m_1 + m_2}\mathbf{r}_{rel} \end{aligned} \tag{6.18}$$

[9] We could of course equally well make $\mathbf{r}_{rel}$ the location of m_1 relative to m_2. It's all relative.

[10] Actually, you could just read the results for $\mathbf{r}_{1,cm}$ and $\mathbf{r}_{2,cm}$ straight off of eqq. (6.17) by noting that in $\mathbf{r}_i = \mathbf{r}_{cm} + \mathbf{r}_{i,cm}$ the $\mathbf{r}_{i,cm}$ are whatever is added to $\mathbf{r}_{cm}$ to get $\mathbf{r}_i$.

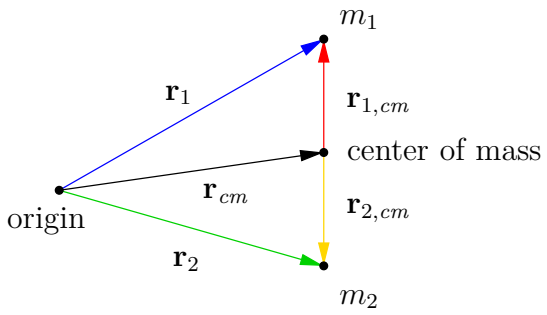


Figure 6.7: Coordinates Relative to the Center of Mass

Taking a time derivative of eqq. (6.18), we obtain corresponding results for the velocities of m_1 and m_2 relative to the center of mass:

$$\begin{aligned} \mathbf{v}_{1,cm} &= -\frac{m_2}{m_1+m_2}\mathbf{v}_{rel} \\ \mathbf{v}_{2,cm} &= \frac{m_1}{m_1+m_2}\mathbf{v}_{rel} \end{aligned} \tag{6.19}$$

Recalling that the center-of-mass frame could equally well be called the center-of-momentum frame, we see that the momenta of m_1 and m_2 in the center-of-mass frame are, as expected, back-to-back, so that the total momentum vanishes:

$$\begin{aligned} m_1\mathbf{v}_{1,cm} &= m_1\left(-\frac{m_2}{m_1+m_2}\mathbf{v}_{rel}\right) = -\frac{m_1m_2}{m_1+m_2}\mathbf{v}_{rel} \\ m_2\mathbf{v}_{2,cm} &= m_2\left(\frac{m_1}{m_1+m_2}\mathbf{v}_{rel}\right) = \frac{m_1m_2}{m_1+m_2}\mathbf{v}_{rel} \\ \mathbf{p}_{total} &= m_1\mathbf{v}_{1,cm} + m_2\mathbf{v}_{2,cm} \\ &= -\frac{m_1m_2}{m_1+m_2}\mathbf{v}_{rel} + \frac{m_1m_2}{m_1+m_2}\mathbf{v}_{rel} = 0 \quad \checkmark \end{aligned}$$

The combination $m_1m_2/(m_1+m_2)$ that we have encountered here occurs frequently enough to have its own notation:

$$\mu = \frac{m_1m_2}{m_1+m_2} \tag{6.20}$$

This animal μ is called the reduced mass because it has the dimensions of a mass and is, as you can see from the fraction by which it is defined, less than either of m_1 or m_2. In terms of μ, the momenta of m_1 and m_2 have the simple form

$$\begin{aligned} m_1\mathbf{v}_{1,cm} &= -\mu\mathbf{v}_{rel} \\ m_2\mathbf{v}_{2,cm} &= \mu\mathbf{v}_{rel} \end{aligned}$$

The expression for the total kinetic energy turns out to be similarly simple. Taking a time derivative of eqq. (6.17), we obtain results for the velocities of m_1 and m_2:

$$\begin{aligned} \mathbf{v}_1 &= \mathbf{v}_{cm} - \frac{m_2}{m_1+m_2}\mathbf{v}_{rel} \\ \mathbf{v}_2 &= \mathbf{v}_{cm} + \frac{m_1}{m_1+m_2}\mathbf{v}_{rel} \end{aligned} \tag{6.21}$$

We can now use these expressions for the velocities of m_1 and m_2 in the total kinetic energy, which works out, when we expand and then combine like terms, to

$$\begin{aligned} K_{total} &= \tfrac{1}{2}m_1 v_1^2 + \tfrac{1}{2}m_2 v_2^2 \\ &= \tfrac{1}{2}m_1\left(\mathbf{v}_{cm} - \frac{m_2}{m_1+m_2}\mathbf{v}_{rel}\right)^2 + \tfrac{1}{2}m_2\left(\mathbf{v}_{cm} + \frac{m_1}{m_1+m_2}\mathbf{v}_{rel}\right)^2 \\ &= \tfrac{1}{2}m_1\left(v_{cm}^2 - 2\frac{m_2}{m_1+m_2}\mathbf{v}_{cm}\cdot\mathbf{v}_{rel} + \left(\frac{m_2}{m_1+m_2}\right)^2 v_{rel}^2\right) \\ &\quad + \tfrac{1}{2}m_2\left(v_{cm}^2 + 2\frac{m_1}{m_1+m_2}\mathbf{v}_{cm}\cdot\mathbf{v}_{rel} + \left(\frac{m_1}{m_1+m_2}\right)^2 v_{rel}^2\right) \\ &= \tfrac{1}{2}(m_1+m_2)v_{cm}^2 + \tfrac{1}{2}\left(m_1\left(\frac{m_2}{m_1+m_2}\right)^2 + m_2\left(\frac{m_1}{m_1+m_2}\right)^2\right)v_{rel}^2 \\ &= \tfrac{1}{2}(m_1+m_2)v_{cm}^2 + \tfrac{1}{2}\left(\frac{m_1 m_2}{m_1+m_2}\right)v_{rel}^2 \\ &= \tfrac{1}{2}Mv_{cm}^2 + \tfrac{1}{2}\mu v_{rel}^2 \end{aligned}$$

So the decomposition of kinetic energy into center-of-mass and relative contributions works out in the two-body case to

$$K_{total} = K_{cm} + K_{rel} \tag{6.22a}$$

where

$$K_{cm} = \tfrac{1}{2}Mv_{cm}^2 \tag{6.22b}$$

and

$$K_{rel} = \tfrac{1}{2}\mu v_{rel}^2 \tag{6.22c}$$

Eqq. (6.22a) and (6.22b) are, as we would have expected, identical to the eqq. (6.13a) and (6.13b) for the general (that is, n-body) case, and we could in fact have obtained eq. (6.22c) simply by using eqq. (6.21) in eq. (6.13c):

$$\begin{aligned} K_{rel} &= \tfrac{1}{2}m_1 v_{1,cm}^2 + \tfrac{1}{2}m_2 v_{2,cm}^2 \\ &= \tfrac{1}{2}m_1\left(-\frac{m_2}{m_1+m_2}\mathbf{v}_{rel}\right)^2 + \tfrac{1}{2}m_2\left(\frac{m_1}{m_1+m_2}\mathbf{v}_{rel}\right)^2 \end{aligned}$$

$$
\begin{aligned}
&= \tfrac{1}{2}\left(m_1 \left(\frac{m_2}{m_1+m_2} \right)^2 + m_2 \left(\frac{m_1}{m_1+m_2} \right)^2 \right) v_{rel}^2 \\
&= \tfrac{1}{2}\left(\frac{m_1 m_2}{m_1+m_2} \right) v_{rel}^2 \\
&= \tfrac{1}{2}\mu v_{rel}^2
\end{aligned}
$$

But we figured the reinforcement of going through the derivation for the two-body case from scratch might be beneficial.

In the center-of-mass frame, where $\mathbf{v}_{cm} = 0$ and thus $K_{cm} = 0$, this total kinetic energy is, as expected, entirely due to the relative motion of m_1 and m_2, and it has the particularly simply expression $\frac{1}{2}\mu v_{rel}^2$. Again, if momentum is conserved, only this relative kinetic energy can change, and in the case of a completely inelastic collision it is entirely lost as the two masses stick together.

Eq. (6.22c) is yet another hugely important result: remembering that the relative kinetic energy in a two-body collision is $\frac{1}{2}\mu v_{rel}^2$ can save you a *lot* of work; what would otherwise be a very involved problem will often become a one- or two-liner.

6.7.1 The One-Dimensional Two-Body Elastic Collision

One special case with a number of practical applications is the one-dimensional elastic collision between two bodies. If we denote the velocities of the two masses m_1 and m_2 by v_1 and v_2 before the collision and by v_1' and v_2' after the collision, then our momentum conservation relation is

$$m_1 v_1 + m_2 v_2 = m_1 v_1' + m_2 v_2' \tag{6.23}$$

Since the collision is elastic, we also have conservation of kinetic energy:

$$\tfrac{1}{2} m_1 v_1^2 + \tfrac{1}{2} m_2 v_2^2 = \tfrac{1}{2} m_1 v_1'^2 + \tfrac{1}{2} m_2 v_2'^2 \tag{6.24}$$

Assuming that we know the masses and the velocities before the collision, then eqq. (6.23) and (6.24) can be solved for the velocities v_1' and v_2' after the collision. This is most easily accomplished by formally solving the momentum equation for one of the v''s and substituting the result into the kinetic energy equation. The algebra is tedious but straightforward, so we won't show it here; because the kinetic energy equation is quadratic, you get two pairs of solutions:

$$
\begin{aligned}
v_1' &= v_1 \\
v_2' &= v_2
\end{aligned}
\tag{6.25a}
$$

and

$$\begin{aligned} v_1' &= \frac{(m_1 - m_2)v_1 + 2m_2 v_2}{m_1 + m_2} \\ v_2' &= \frac{(m_2 - m_1)v_2 + 2m_1 v_1}{m_1 + m_2} \end{aligned} \tag{6.25b}$$

Note that both sets of solutions in (6.25a) and (6.25b) are symmetric under the interchange of the subscripts 1 and 2: since the original eqq. (6.23) and (6.24) had this symmetry (which corresponds to it not mattering which mass you call m_1 and which m_2), so must also the solutions to these equations.

Solutions (6.25a), which obviously satisfy eqq. (6.23) and (6.24), correspond to the case where the velocities v_1 and v_2 are such that no collision even occurs: if v_1 and v_2 were such that there were a collision, to have the same velocities after the collision would mean that the masses had somehow passed through each other like ghosts.

The solutions we want, the ones you get when a collision does in fact occur, are (6.25b). In the above general form, they are not terribly illuminating, but there are several special cases of interest:

Billiard balls. If, as is the case with billiard balls, the masses are equal, then $m_1 = m_2 = m$ and eqq. (6.25b) reduce to

$$\begin{aligned} v_1' &= \frac{(m - m)v_1 + 2m v_2}{m + m} = v_2 \\ v_2' &= \frac{(m - m)v_2 + 2m v_1}{m + m} = v_1 \end{aligned}$$

That is, the balls simply exchange velocities. Though collisions between billiard balls are generally two-dimensional, a head-on collision will be one-dimensional. The most commonly occurring such case is when the cue ball strikes a stationary ball dead center: the cue ball then stops, and the struck ball moves off with the cue ball's original velocity.[11]

You'll shoot your eye out! Only a complete moron would fire a BB gun straight into a brick wall. In this case, the mass m_1 of the BB is tiny compared to the mass m_2 of the brick wall ($m_1 \ll m_2$), so that we can neglect it (that is, set it effectively to zero) by comparison to m_2. Also, the wall is stationary, so that $v_2 = 0$. In this case, solutions (6.25b) become

$$v_1' \approx \frac{(0 - m_2)v_1 + 2m_2(0)}{0 + m_2} = -v_1$$

[11] This assumes that there are no complications from backspin, topspin, or "English" on the cue ball, which unfortunately there almost always are. But you nonetheless often see this simple effect in some approximation.

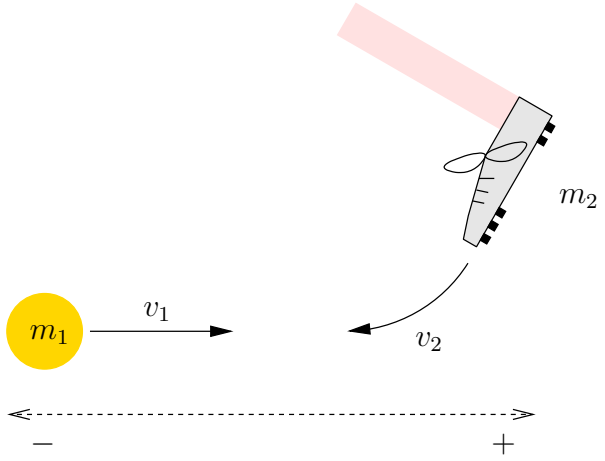


Figure 6.8: Kicking a Soccer Ball

$$v_2' \approx \frac{(m_2 - 0)(0) + 2(0)v_1}{0 + m_2} = 0$$

That is, the wall doesn't acquire any significant velocity as a result of the collision, and the BB pretty much reverses its own velocity and comes straight back at you.

Kicking a soccer ball. Fig. (6.8) is an admittedly rather abstract representation of what happens when you kick a soccer ball that is rolling toward you. Note that we have set up our directions so that the ball is moving in the positive direction and your foot in the negative direction (v_1 is positive and v_2 negative). In this case, your foot and leg are much more massive than the ball, so that again $m_1 \ll m_2$ and we can neglect m_1 by comparison to m_2. Solutions (6.25b) now become

$$v_1' \approx \frac{(0 - m_2)v_1 + 2m_2v_2}{0 + m_2} = -v_1 + 2v_2$$
$$v_2' \approx \frac{(m_2 - 0)v_2 + 2(0)v_1}{0 + m_2} = v_2$$

That is, your leg continues on its merry way without any significant change in velocity, while the ball does two things:

- Reverses its own velocity (the $-v_1$).
- Picks up twice your foot's velocity (the $2v_2$).

Since v_1 was positive and v_2 negative, both of these contributions to v_1' are in the negative direction, corresponding to your getting off really hard shots on rebounds or other situations where the ball was coming toward you. This

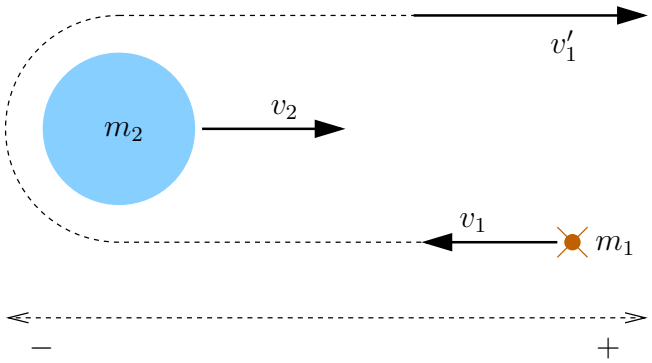


Figure 6.9: The Gravitational Slingshot

same situation occurs in many other games, including baseball, where the pitched ball reverses its own velocity and picks up twice the speed of the bat — which is why you can hit balls a ton farther than you can throw them.

For cases where the ball is rolling away from you as you kick it, v_1 above would have instead a negative value, so that $-v_1$ would be in the positive direction and work against the $2v_2$ contribution from your kick by partially canceling it out. This corresponds to your getting off relatively wimpy shots when you have to strike at balls that are rolling away from you.

Ye olde gravitational slingshot. Consider a spacecraft that gets whipped around by a planet's gravity as the planet whizzes along with its own orbital velocity v_2, as shown in fig.(6.9). As in the preceding case, $m_1 \ll m_2$, so that we get the same solutions:

$$v_1' \approx -v_1 + 2v_2$$
$$v_2' \approx v_2$$

The spacecraft thus reverses its incoming velocity and, in addition, picks up twice the orbital velocity of the planet. Since planets move at tremendous orbital velocities, this is a very effective way to boost the speed of spacecraft.

6.8 Summary of Important Points

Just so you don't lose sight of the many important points in the physics of this chapter:

- For a one-dimensional collection of discrete point masses the location of the center of mass is given by

$$x_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i x_i$$

where M is the total mass. In higher dimensions, this relation becomes

$$\mathbf{r}_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i \mathbf{r}_i$$

For continuous distributions of mass,

$$\mathbf{r}_{cm} = \frac{1}{M}\int_{\text{distribution}} dm\, \mathbf{r}$$

with the total mass of the distribution given by

$$M = \int_{\text{distribution}} dm$$

where

$$dm = \begin{cases} \lambda\, ds & \text{for a linear distribution} \\ \sigma\, dA & \text{for a surface distribution} \\ \rho\, dV & \text{for a volume distribution} \end{cases}$$

The expressions for ds, dA, and dV depend on the geometry of the distribution and can be found in Chapter 2 for Cartesian, cylindrical, and spherical coordinates.

- The center of mass of aggregate distributions of mass can be found by applying eq. (6.2) to the various parts of the distribution, with the sum carried out over the parts of the distribution, each term in the sum consisting of the product of that part's mass $m_{part,i}$ with the location $\mathbf{r}_{part,i}$ of its center of mass:

$$\mathbf{r}_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_{part,i}\, \mathbf{r}_{part,i}$$

Holes can be treated as contributions of negative mass.

- The velocity and acceleration of the center of mass are given by similar weighted averages:

$$\mathbf{v}_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i \mathbf{v}_i \qquad\qquad \mathbf{a}_{cm} = \frac{1}{M}\sum_{i=1}^{n} m_i \mathbf{a}_i$$

- The net external force on a system determines the acceleration of the system's center of mass:

$$\mathbf{F}_{net,\ ext} = M\mathbf{a}_{cm}$$

- With momentum defined by $\mathbf{p} = m\mathbf{v}$ Newton's second law can be expressed as

$$\mathbf{F} = \frac{d\mathbf{p}}{dt}$$

For a system, this becomes

$$\mathbf{F}_{net\ ext,\ system} = \frac{d\mathbf{p}_{total,\ system}}{dt}$$

When no net external force acts on a system, its momentum is exactly conserved. Even if there is a net external force on the system, its momentum is still approximately conserved when a collision occurs quickly enough that that net external force does not have time to significantly change the system's momentum. In this latter case, the total momentum of the system immediately before the collision can be approximately equated to its total momentum immediately after the collision.

- Even if the total momentum of a system is conserved in a collision, in general its total kinetic energy is not. When kinetic energy is not conserved, the collision is termed inelastic. In the special case that kinetic energy is conserved, the collision is termed elastic.

- The total momentum of a system can be expressed as

$$\mathbf{p}_{total,\ system} = M\mathbf{v}_{cm}$$

The total kinetic energy of a system can similarly be expressed as

$$K_{total} = K_{cm} + K_{rel}$$

with the kinetic energy K_{cm} of the center of mass and the kinetic energy K_{rel} relative to the center of mass given by

$$K_{cm} = \tfrac{1}{2}Mv_{cm}^2 \qquad\qquad K_{rel} = \sum_{i=1}^{n} \tfrac{1}{2}m_i v_{i,cm}^2$$

where $v_{i,cm}$ is the velocity of mass m_i relative to the center of mass. In the two-body case, the relative kinetic energy reduces to

$$K_{rel} = \tfrac{1}{2}\mu v_{rel}^2$$

where v_{rel} is the relative velocity of the two bodies and where the reduced mass μ is given by

$$\mu = \frac{m_1 m_2}{m_1 + m_2}$$

- When there is no net external force on a system,

$$\mathbf{p}_{total,\ system} = M\mathbf{v}_{cm} = \text{const}$$

means that the velocity $\mathbf{v}_{cm}$ of the center of mass, and hence also its kinetic energy K_{cm}, are constant. Thus only the relative contribution K_{rel} to the kinetic energy can change. Since $K_{rel} \geq 0$, there is an upper limit to how much kinetic energy can be lost in a collision: the worst possible case is a completely inelastic collision, in which everything sticks together, so that there is no relative motion after the collision.[12]

- The general result for the one-dimensional two-body elastic collision is

$$v_1' = \frac{(m_1 - m_2)v_1 + 2m_2v_2}{m_1 + m_2}$$
$$v_2' = \frac{(m_2 - m_1)v_2 + 2m_1v_1}{m_1 + m_2}$$

You should *not* memorize this result, but you should be able to work it out from momentum and kinetic energy conservation — at least for special cases in which the algebra is relatively simple. You should also be able to apply and interpret this result in various limits of the masses.

6.9 Some Gravitational Yawing

There are a few important aspects of gravitational forces and energies that have to be covered somewhere in the course but that neither merit their own chapter nor fit neatly into any existing chapter. Since the math and techniques involved in some of these miscellaneous aspects of gravity are similar to those used in center-of-mass calculations, we are, for better or worse, plunking them down here.

It turns out that a uniform thin[13] spherical shell of mass m and radius a has two properties:

- From the outside (that is, for $r > a$), the shell is gravitationally equivalent to a point mass m located at the center of the shell. In other words, it is as though all of the shell's mass were concentrated at its center, so that the gravitational force that would be exerted on a point mass m' a distance r from the shell's center is

$$F = \frac{Gmm'}{r^2}$$

just as it would be for two point masses.

[12] Again, as noted in footnote 8 on p.303, this is ignoring rotational contributions to the system's kinetic energy, a complication with which we will not deal until Chapter 7.

[13] *Thin* meaning the shell is essentially of zero thickness.

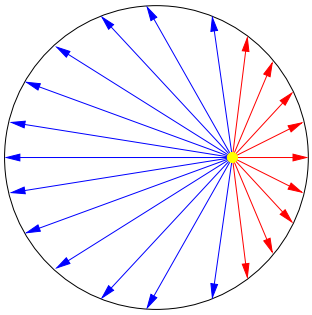


Figure 6.10: Inside of Uniform Spherical Shell

- On the inside (that is, for $r < a$), there is no net gravitational force due to the shell. It is as though the shell isn't even there, so that no gravitational force would be exerted on a point mass m' located anywhere within the interior of the shell.

We could prove these properties by doing an integration over a spherical shell of uniform mass density, but the integration, which turns out to be a moderately serious pain, would be unlikely to leave you a better person for having been through the experience. We will therefore simply accept these properties of the gravitational force for now and postpone their proof until p.784, at which point Gauss's theorem and law will make them much easier to prove. But the equivalence of the shell to a point mass for points outside its radius should seem plausible, and you can make sense of the vanishing of the gravitational force on the interior of the shell from fig. (6.10): The yellow dot in the figure represents a point mass on the interior of the shell, somewhere to the right of the shell's center. As you can see from the red arrows, those points on the shell to the right of the point mass will, by symmetry, exert on the point mass a net gravitational force directly to the right. And from the blue arrows you can see that those points on the shell to the left of the point mass will, by the same symmetry, exert on the point mass a net gravitational force directly to the left. Fewer points on the shell lie to the right of the point mass than to the left, but those points to the right of the point mass are closer to the point mass and therefore exert a stronger gravitational force on it. It is by no means obvious, but it does turn out that a sphere is just the right shape to balance these competing effects exactly, with the result that there is no net gravitational force on the point mass.

Note that these properties of a shell of mass depend critically both on the shell being spherical and on the shell's mass being uniformly distributed over the shell; if the shell is not spherical or the mass distribution is uneven, neither property holds.

Since solid spheres of mass or spherical shells of finite thickness can be regarded as collections of thin spherical shells, the above properties also apply

to uniform solid spheres and uniform spherical shells of finite thickness, albeit in a slightly different way for their interiors:

- From the outside, the solid sphere or shell is gravitationally equivalent to a point mass located at its center.
- On the inside, if a point mass lies a distance r' from the center, only that part of the sphere's or shell's mass interior to r' (that is, that fraction of the sphere's or shell's mass that lies at radii $r < r'$) contributes to the net gravitational force on the point mass.

Note also that while the distribution of mass within the sphere or shell has to be rotationally uniform it can be a function of radius: a density that depends on just r (as opposed to also depending on the spherical angles θ or φ) will still constitute a collection of thin spherical shells, the mass of each of which is uniformly distributed over its surface.

And with that we have arrived at the point of showing you how you make use of all of this. As an example, consider a solid sphere of radius a, the density ρ of which falls off as $1/r^2$:

$$\rho = \rho_0 \frac{a^2}{r^2}$$

where ρ_0 is a constant. To relate ρ_0 to the total mass m of the sphere, we simply integrate the density over the volume of the sphere:

$$\begin{aligned} m &= \int\limits_{\text{sphere}} dV\, \rho \\ &= \int\limits_{\text{sphere}} r^2 \sin\theta\, dr\, d\theta\, d\varphi\, \rho_0 \frac{a^2}{r^2} \\ &= \rho_0 a^2 \int_0^a dr \int_0^\pi \sin\theta\, d\theta \int_0^{2\pi} d\varphi \\ &= \rho_0 a^2 (a) \Big[-\cos\theta \Big]_0^\pi (2\pi) \\ &= \rho_0 a^2 (a)(2)(2\pi) \\ &= 4\pi \rho_0 a^3 \end{aligned}$$

Thus $\rho_0 = m/4\pi a^3$ and we may write the density as

$$\rho = \frac{m}{4\pi a^3} \frac{a^2}{r^2} = \frac{m}{4\pi a r^2}$$

If now we want to calculate the force that would be exerted by this sphere on a point mass m' at any radius from $r = 0$ to $r = \infty$, we first note that

outside the sphere the force is the same as if the total mass m of the sphere were collapsed down to a point mass at its center:

$$F = \frac{Gmm'}{r^2} \qquad (r \geq a) \tag{6.26}$$

When m' is inside the sphere, only that fraction of the sphere's mass that is interior to the radius of m' contributes to the net force, so we need to calculate that mass fraction by carrying out an integration otherwise identical to the one we did above, but out to a general radius $r \leq a$ rather than all the way out to $r = a$:

$$\begin{aligned} m_{interior} &= \rho_0 a^2 \int_0^r dr \int_0^\pi \sin\theta \, d\theta \int_0^{2\pi} d\varphi \\ &= \rho_0 a^2 (r)(2)(2\pi) \\ &= \frac{m}{4\pi a^3} \, a^2 (r)(2)(2\pi) \\ &= m\frac{r}{a} \end{aligned}$$

The force exerted on m' at points on the interior of the sphere is then the same as if this mass mr/a were collapsed down to a point mass at its center: [14]

$$F = \frac{Gm_{interior}m'}{r^2} = \frac{G\left(m\frac{r}{a}\right)m'}{r^2} = \frac{Gmm'}{ar} \qquad (r \leq a) \tag{6.27}$$

Now suppose we want to calculate the gravitational potential energy of m', with the potential taken to be zero at $r = \infty$. If we come straight in along the radial direction from $r = \infty$ and remember that the force on m' is inward, that is, in the negative radial direction, for points outside the sphere we have

$$\begin{aligned} U &= -\int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{F} \cdot d\mathbf{r} \\ &= -\int_\infty^r \left(-\frac{Gmm'}{r^2}\right) dr \\ &= -\left[\frac{Gmm'}{r}\right]_\infty^r \end{aligned}$$

[14] We are being a bit glib about the special case $r = a$. When there is a finite amount of mass in an infinitesimally thin shell of radius a, then there is a finite discontinuity in the gravitational force as we pass from $r > a$ to $r < a$. But in cases like the present example, when the infinitesimal shells that make up the distribution have correspondingly infinitesimal masses, there is no discontinuity to worry about, and the gravitational force is the same for $r \to a^+$ and $r \to a^-$.

$$= -\frac{Gmm'}{r} \qquad (r \geq a)$$

as we would have expected: since the force is the same that we would have for two point masses, so is the potential energy. For points on the interior of the sphere, however, we have to be more careful: when carrying out the integration for the potential energy from $r = \infty$ to a radius $r < a$, we have to remember that the force is given by eq. (6.26) when $r \geq a$ and by eq. (6.27) when $r \leq a$ and split up our integration for the potential energy accordingly:

$$\begin{aligned}
U &= -\int_{r=\infty}^{r} \mathbf{F} \cdot d\mathbf{r} \\
&= -\int_{r=\infty}^{a} \mathbf{F} \cdot d\mathbf{r} - \int_{r=a}^{r} \mathbf{F} \cdot d\mathbf{r} \\
&= -\int_{r=\infty}^{a} \left(-\frac{Gmm'}{r^2}\right) dr - \int_{r=a}^{r} \left(-\frac{Gmm'}{ar}\right) dr \\
&= -\left[\frac{Gmm'}{r}\right]_{\infty}^{a} - \left[\frac{Gmm'}{a} \ln r\right]_{a}^{r} \\
&= -\frac{Gmm'}{a} - \frac{Gmm'}{a} \ln \frac{r}{a} \\
&= -\frac{Gmm'}{a}\left(1 + \ln \frac{r}{a}\right) \qquad (r \leq a)
\end{aligned}$$

As a second example that illustrates some other, unrelated aspects of gravitational forces and energies, consider the uniform semicircular arc of mass m and radius a shown in fig. (6.11). We can calculate the net gravitational force on a point mass m' at the "center" of the arc by regarding the arc as a collection of infinitesimal mass points dm, each mass point corresponding to an infinitesimal bit of arc ds that makes up the semicircle. Since the mass m is uniformly distributed over the length πa of the arc, we have

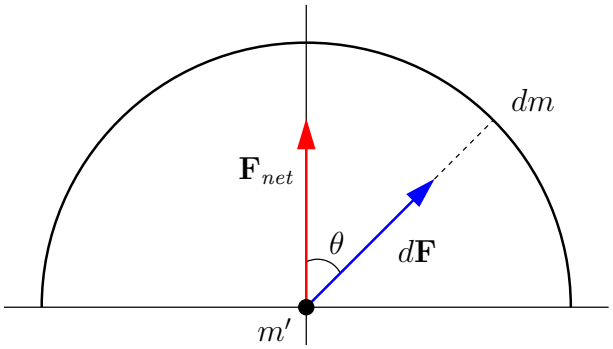


Figure 6.11: The Old Gravitational-Arc Trick

$\lambda = m/\pi a$ for the linear mass density of the arc and

$$dm = \lambda\, ds = \frac{m}{\pi a}\, ds$$

The magnitude of the contribution of dm to the net force on m' (the blue arrow in fig. (6.11)) is thus

$$dF = \frac{Gm'\, dm}{a^2} = \frac{Gm'\left(\frac{m}{\pi a}\, ds\right)}{a^2} = \frac{Gm'm}{\pi a^3}\, ds$$

The complication is that to get the net force on m' we need to take the *vector* sum of all these contributions dF. But from the symmetry of the arc we know that the net force on m' must be along the bisector of the arc, as shown by the red arrow in fig. (6.11). We can therefore get the contribution of dF to the net force by the expedient of including a trig factor: only the component of $d\mathbf{F}$ along $\mathbf{F}_{net}$ will contribute to the net force, and this component is given by

$$dF\cos\theta = \frac{Gm'm}{\pi a^3}\, ds\, \cos\theta$$

Thus

$$F_{net} = \int_{\text{arc}} dF\cos\theta = \int_{\text{arc}} \frac{Gm'm}{\pi a^3}\, ds\, \cos\theta = \frac{Gm'm}{\pi a^3}\int_{\text{arc}} ds\, \cos\theta$$

To carry out this integration, we need to express ds in terms of θ: $ds = a\, d\theta$, with the arc corresponding to $-\frac{\pi}{2} \le \theta \le \frac{\pi}{2}$. We therefore have

$$\begin{aligned} F_{net} &= \frac{Gm'm}{\pi a^3}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} a\, d\theta\, \cos\theta \\ &= \frac{Gm'm}{\pi a^3}\, a\Big[\sin\theta\Big]_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \\ &= \frac{Gm'm}{\pi a^3}\, a(2) \\ &= \frac{2Gm'm}{\pi a^2} \end{aligned}$$

If we want to calculate the gravitational potential energy of m', with the potential taken to be zero at $r = \infty$, we can again regard the arc as a collection of infinitesimal mass points dm. The contribution dU to the potential energy of m' due to the mass dm on each bit of arc will be given by the usual expression for the potential energy of two mass points:

$$dU = -\frac{Gm'\, dm}{a}$$

We could again express dm in terms of ds and hence $d\theta$ in order to integrate these contributions over the arc, but since potential energy is a scalar and not a vector, there is no trig factor to take into account and we don't need to make things so complicated: when we integrate over the contributions dU we have simply

$$U = \int dU = -\int \frac{Gm'\,dm}{a} = -\frac{Gm'}{a}\int dm = -\frac{Gm'm}{a}$$

Because all of the bits of mass that make up the arc are the same distance a from m', this result potential energy is the same as we would have for two point masses separated by a distance a.[15]

And with that our little adventure in miscellaneous aspects of gravitational forces and energies is concluded.

[15] We could in fact have made this observation and immediately written down the result for the potential energy, but the whole point of the example was to illustrate the general technique of integrating over scalar contributions dU to get the total potential energy.

6.10 Problems

1. Because it somehow just seemed like the thing to do at the time, you took a meter stick of mass m and placed point-like masses of $\frac{1}{2}m$ and $\frac{3}{2}m$ at its 80- and 30-cm marks, respectively.

   (a) At what point would the arrangement have been balanced if you had placed a finger as a fulcrum under that point? (For reasons that will become clear when we get to torques and static equilibria in Chapters 7 and 8, this balance point, known as the center of gravity, is at the center of mass.)

   (b) If you had kept the mass $\frac{1}{2}m$ at the 80-cm mark and wanted the balance point to be at the center of the stick, where should the mass $\frac{3}{2}m$ have instead been placed?

2. At a summer camp for children who lack imagination and aptitude, you make an ashtray in pottery class, even though no one in your family smokes. Which is just as well, because the ashtray subsequently explodes in the kiln, anyway, taking with it a half dozen surrounding ashtrays made by your fellow campers. You then make a chessboard in woodworking class, even though no one in your family plays chess. Which could be why you make all 64 squares perfectly uniform, including their color. Anyway, if the side of each of the squares has length ℓ and the craftsmanship is such that the Q4 and K4 squares (d4 and e4 for those of you who use algebraic notation, and just a complete mystery to those of you who don't know anything about chess) immediately fall out, where is the center of mass of what is left of the board?

3. At a pagan ritual in the basement, Gaffney carefully draws a circle of radius a, inscribes in it a regular polygon of 37 sides, places a candle of mass m at 36 of the 37 vertices, and then bites the head off of a chicken. Where is the center of mass of the candles relative to the center of the circle?

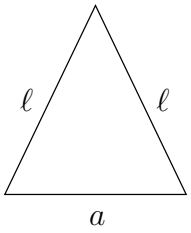


Figure 6.12: Problem 4

4. For no particular reason, an isosceles triangle with the dimensions shown in fig. (6.12) is constructed out of wire of uniform linear density λ.

   (a) Determine the location of the triangle's center of mass. Note that no integrations are necessary.

   (b) Explain how and why your result does or does not make sense in the limit of

      i. Large ℓ.

      ii. Small a.

      iii. Small ℓ.

5. An empty soup can (a right circular cylinder of radius a and height ℓ, with a top and bottom) is cut in half, so that you end up with a height of $\frac{1}{2}\ell$ and a bottom but no top. Locate the center of mass of the resulting "half can."

6. A thin rod of length ℓ has a weirdly biased mass distribution: its linear mass density is proportional to the distance from its lighter end, so that $\lambda = \alpha x$, where x is the distance from the lighter end of the rod and α is a constant.

   (a) What are the physical dimensions of α?

   (b) By integrating over the rod, obtain a result for the mass m of the rod.

   (c) Determine the location of the center of mass of the rod.

7. Consider the triangle formed by the x axis and the lines $x = \ell$ and $y = \alpha x$.

   (a) If the triangle's mass m is evenly spread over its area, what is its mass density?

   (b) Determine the location of the center of mass of the triangle.

   (c) Make physical sense of your result for the location of the center of mass in the limit $\alpha \to 1$.

   (d) Explain how the x component of your result for the location of the center of mass reproduces your result for #6c.

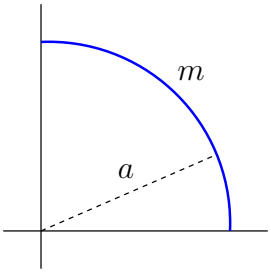


Figure 6.13: Problem 8

8. Fig. (6.13) shows an arc consisting of one quarter of a circle of radius a. A mass m is evenly distributed along this arc. Not very exciting, but there you have it.

   (a) What is the linear mass density of the arc?

   (b) Determine the x coordinate of the center of mass of the arc. See the footnote if you need a hint.[16]

   (c) Determine the y coordinate of the center of mass of the arc.

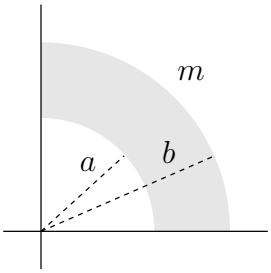


Figure 6.14: Problem 9

9. Fig. (6.14) shows an equally unexciting quarter-annulus (that is, a quarter of a washer) of inner radius a and outer radius b, over which a mass m is evenly distributed.

   (a) What is the surface mass density of the quarter-annulus?

   (b) Determine the x and y coordinates of the center of mass of the quarter-annulus

   (c) Show that in the limit $b \to a$ this reproduces your result for #8.

[16] $dm = \lambda\, ds$, where $ds = a\, d\theta$.

10. Consider the first octant of a spherical surface of radius a. Exciting, eh?

    (a) What is the surface mass density of this animal?

    (b) Determine the x, y, and z coordinates of the center of mass of this animal.

11. You (mass m) walk from one extreme end of a boat (mass m', length ℓ) to the other.

    (a) If there is no friction with the water, how far does the boat shift relative to the shore, and in what direction? See the footnote if you need a hint.[17]

    (b) Make physical sense of your result for

        i. $m \gg m'$.

        ii. $m' \gg m$.

        iii. $m = m'$.

    (c) In reality, there is friction with the water. How will this friction affect how far the boat shifts? Will it make a difference whether you walk to the other end of the boat quickly or slowly?

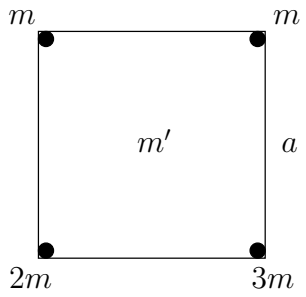


Figure 6.15: Problem 12

12. Four people, whose masses are as shown in fig. (6.15), are at the extreme corners of a square raft of side a and mass m' that is afloat on conveniently frictionless water. Just to make your life difficult, the people diagonally opposite each other switch places.

    How far and in what direction does the raft shift relative to the shore? This problem is not bad at all if you keep your wits about you; you should already be aware, from #11, of the physical condition that applies.

[17] Think about how the center of mass of the system moves. Also, although it might at first seem otherwise, you do not need to know where the center of mass of the boat is to obtain a result.

13. *Quantitatively* describe the motion of the center of mass (that is, the location, velocity, and acceleration of the center of mass) of the two masses in Atwood's machine, assuming that the masses start from rest at the same height (that is, lined up vertically with each other). (See p.231 if you've forgotten what an Atwood's machine is.)

14. Two totally uninspiring masses, m_1 and $m_2 > m_1$, move at speeds v_1 and v_2, respectively.

    (a) If $p_1 = p_2$, what is the relationship between their kinetic energies?

    (b) If $K_1 = K_2$, what is the relationship between their momenta?

15. (Debunking Hollywood.) Dirty Harry (80 kg) casually walks down the street chewing on a hamburger as he blows people away right and left with a .44 magnum (bullet mass 240 grains = 15.5 g, muzzle velocity 440 m/s).

    (a) One 80 kg bad guy is standing at rest as he is shot. If the bullet sticks in him, what backward velocity does he acquire as a result of being shot?

    (b) Another 80 kg bad guy charges Dirty Harry at 4.0 m/s as he is shot. If the bullet sticks in him as well, how much is he slowed down?

    (c) Hollywood would have you believe that such bad guys always go flying backward through a plate-glass window or a pile of garbage cans as a result of being shot. Your answers to the preceding parts of the problem should have debunked that illusion, but, if it *were* true, what consequences would conservation of momentum then have for Dirty Harry and the recoil he experienced?

16. While playing soccer, Gaffney (mass m) and an opponent (mass $2m$) each run at speed v toward a 50-50 ball.

    (a) i. If their collision were elastic, what velocities would they have after the collision? Work this out from scratch; don't cheat by going back to look at the result for the general case.

    ii. What, if anything, can you conclude about the nature of the force that Gaffney and the opponent exert on each other during this collision?

    (b) i. Instead, Gaffney is knocked flat on his butt. What is the speed of his opponent after the collision, how much kinetic energy is lost in the collision, and where does this lost kinetic energy go?

    ii. What, if anything, can you conclude about the nature of the force that Gaffney and the opponent exert on each other during this collision?

17. A pink Cadillac (mass $5m$, \$49,999.$\underline{95}$ including tax) with heart-shaped windows, leopard-skin upholstery, fuzzy dice hanging from the mirror, a faux-fur steering-wheel cover, one of those little cardboard pine-scented Christmas trees on the dashboard, and a bobble-head of Elvis in the back window, plows into a parked Yugo (mass m, \$49.$\underline{95}$) at speed v.

    (a) If the two cars stick together as a result of the collision,

        i. What is their speed after the collision?

        ii. How much kinetic energy is lost in the collision, and where does it go?

    (b) Suppose instead the two cars collide elastically. What is the velocity of each car after the collision, and in what direction is it moving? Work this out from scratch; don't cheat by going back to look at the result for the general case.

    (c) Repeat the preceding two parts for the case that it is the Yugo that slams into the parked Cadillac at speed v. If you have used good technique in solving the preceding parts of the problem, this shouldn't entail all *that* much work.

    (d) Explain why it is not possible to determine the velocities of the Cadillac and Yugo after the collision if all you know (in addition to the masses and initial velocities) is that the collision is inelastic (not completely inelastic, but simply inelastic).

18. As you stand about to make a catch during a friendly pick-up game of ultimate Frisbee you (mass m) are viciously taken out from behind by a knuckle-dragging, mouth-breathing football player (mass $3m$) moving at speed V. The two of you go down as a single mass and skid a distance ℓ along the field as you come to rest. What was the (constant) force of friction between the two of you and the field?

19. You challenge the football player of #18 to a duel, proposing that a loaded shotgun be placed on the ice of the hockey rink midway between the two of you, and that each of you, starting from rest and taking hold of one end of a rope as in tug-of-war, draw yourself toward the shotgun by pulling on the rope. The football player, who, like most football players, is brashly confident of prevailing by a combination of brute strength and bad attitude, smugly accepts your challenge. Explain physically how this turns out to be one of those rare cases where justice is done.

20. (Déjà vu all over again.) Recall that in #13a on p.226 you were kidnapped by vicious Bokononist space aliens who took you to a barren distant planet, forced you to put on a bunny suit, and left you at rest in the center of a perfectly frictionless and level frozen pond of ice-nine. Analyze the means by which you can get off the pond and escape, this time in terms of momentum rather than forces.

21. Very high-powered rifles, such as the .458 magnum, often come with rubbery pads on the butts of the stocks (the part that you hold against your shoulder) to cushion the recoil.

    (a) Explain, in terms of momentum and force, how these pads help.

    (b) How does this also help explain why the kick from semiautomatic rifles and handguns is not as sharp as that from revolvers and bolt-action rifles?

    (c) Cushioning is ineffective if it is either too soft or too hard. Explain what happens in these two cases.

22. Consider a one-dimensional collision between a mass m, initially moving at velocity v_0, and a stationary mass $2m$. For each of the three cases specified below, determine each of the following by the method indicated:

    (a) The velocities of the masses after the collision, by conservation of momentum

    (b) The velocity of the center of mass, from the total momentum

    (c) The velocity of the center of mass before the collision, by taking a weighted average of the velocities of the two masses before the collision

    (d) The velocity of the center of mass after the collision, by taking a weighted average of the velocities of the two masses after the collision

    (e) The relative velocity of the masses before the collision

    (f) The relative velocity of the masses after the collision

    (g) The total kinetic energy before the collision, by taking the naive sum of the kinetic energies of the individual masses before the collision

    (h) The total kinetic energy after the collision, by taking the naive sum of the kinetic energies of the individual masses after the collision

    (i) The center-of-mass kinetic energy, both before and after the collision, from the velocity of the center of mass

    (j) The relative kinetic energy before the collision, from the relative velocity of the masses before the collision

    (k) The relative kinetic energy after the collision, from the relative velocity of the masses after the collision

    (l) The changes in the total, center-of-mass, and relative kinetic energies

    when

    I. the collision is completely inelastic.

    II. the collision is elastic.

    III. the mass m is at rest after the collision.

    Reflect on the physical significance of your answers and try to find as many correlations and consistencies among them as you can. And don't be a slacker: while this exercise is admittedly rather tedious and pedestrian, it will help solidify your overall understanding of the kinematics of collisions.

23. We have established that internal forces do not contribute to the net force on a system and therefore do not change its total momentum. Explain whether internal forces can or cannot change each of the following, and if they can, give a simple example of a system and an internal force for which they do.

    (a) The total kinetic energy of a system.

    (b) The kinetic energy of the center of mass of a system (in the sense of the K_{cm} of eq. (6.13a)).

    (c) The relative kinetic energy of a system (in the sense of the K_{rel} of eq. (6.13a)).

24. (a) As seen by a stationary observer, two masses, m_1 and m_2, are moving along an x axis with velocities v_1 and v_2, respectively, when they collide. From the perspective of this stationary observer, momentum is conserved and the change in kinetic energy is ΔK. Show that from the perspective of an observer moving down the x axis at velocity u, momentum is still conserved and the change in kinetic energy is the same as that reported by the stationary observer. See the footnote if you need a hint.[18]

    (b) Now extend your proof to the case of a three-dimensional collision among any number of bodies.

25. At a 4th of July celebration, a baby of mass m is wrapped in stars-and-stripes swaddling clothes and loaded into a horizontal spring gun of spring constant k mounted rigidly on a boat afloat in (you guessed it) frictionless water. The spring gun and boat have a total mass m'.

    (a) From the perspective of the spectators standing on the shore, initially everything is at rest and the baby leaves the gun with velocity v when it is fired. How far was the spring compressed when the gun was loaded?

    (b) Suppose now that while the spectators still see the baby leave the gun with velocity v, everything had instead been moving at a common velocity v_0 before the gun was fired. How far was the spring compressed when the gun was loaded?

[18] Make use of eq. (6.13a).

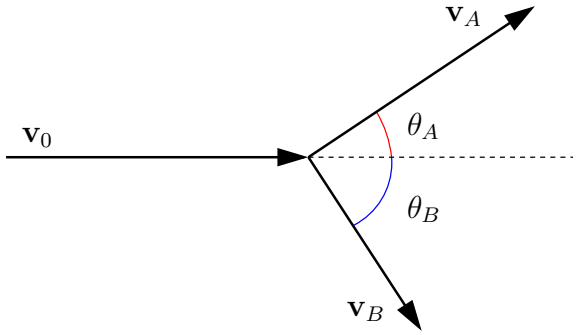


Figure 6.16: Problem 26

26. A billiard ball A moving at speed $\mathbf{v}_0$ strikes another billiard ball B, initially at rest. After the collision, the two balls move off in different directions, as shown in a view from above in fig. (6.16).

    (a) Set up the vector relation for conservation of momentum, leaving the velocities expressed as vectors (that is, without breaking them into components).

    (b) Represent this vector relation graphically as a triangle. Draw and label the velocity vectors in this triangle.

    (c) Because billiard balls are very hard and rigid, their collisions are very nearly elastic. Set up the relation for conserving kinetic energy.

    (d) What do your two relations tell you about the geometry of the collision?

27. At 8:08 one morning you (mass m) are rushing to class at speed $2v$ when you collide at a right angle with a feckless oaf (mass $2m$) lumbering to a different class at speed v.

    (a) If the collision is completely inelastic, at what speed and in what direction are you and the oaf moving after the collision?

    (b) If the collision is elastic, is it possible to solve for the velocities of yourself and the oaf after the collision? If so, set up the necessary relations. If not, explain why not.

28. Of course you would never actually do such a thing, but suppose you and a friend steal a golf cart and go joy-riding around campus at 3:00am on the night of the senior prank. Further suppose that you are challenged by another golf cart loaded with six people, all hurling insults and making rude gestures at you. It being a matter of principle to defend your honor — and, as we all know, senior prank is always a matter of high principle — , you size up the other cart's trajectory and steer toward a collision. At the moment of impact, your cart (total mass $m_{righteous}$, or m_r for short) is moving at speed v_{r0}, and the other cart (total mass m_{losers}, or m_ℓ for short) is moving at a right angle to you at speed $v_{\ell 0}$.

    (a) Set up, but do not solve, the equations needed to completely determine the velocities of the carts after the collision if

        i. The collision is completely inelastic.

        ii. The collision is elastic.

        In each case, note whether these equations are sufficient to determine the velocities after the collision.

    (b) If the collision is elastic, is it possible for your cart to come to rest as a result of the collision? Justify your assertion.

    (c) If the collision is inelastic, but not completely inelastic, is it possible for your cart to come to rest as a result of the collision? Justify your assertion.

29. (a) Show, by arguing from momentum and energy considerations, that when a point mass m with initial velocity

$$\mathbf{v}_0 = v_{0x}\,\hat{\mathbf{x}} + v_{0y}\,\hat{\mathbf{y}} + v_{0z}\,\hat{\mathbf{z}}$$

    collides elastically with an infinitely massive, frictionless wall lying in the xy plane, the point mass will rebound with velocity

$$\mathbf{v} = v_{0x}\,\hat{\mathbf{x}} + v_{0y}\,\hat{\mathbf{y}} - v_{0z}\,\hat{\mathbf{z}}$$

    (b) What impulse is delivered to the point mass? (*Impulse* means simply *change in momentum*, that is, $\Delta\mathbf{p}$. It is therefore a totally unnecessary and utterly silly term; we introduce you to it here just in case you someday encounter out in the world, where people do unnecessary and utterly silly things. In fact, introducing you to this term is really the only point of this part of the problem.)

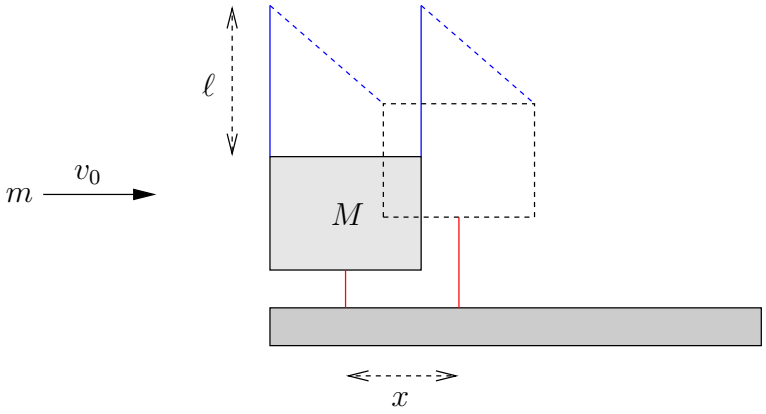


Figure 6.17: Problem 30

30. Fig. (6.17) shows a classic device for determining the speed of high-velocity projectiles like bullets: the ballistic pendulum. The projectile (mass m, velocity v_0) is fired horizontally into the pendulum, which consists of a heavy block (mass M) suspended from cords of length ℓ (the blue cords in fig. (6.17)). As a result of the collision, in which the projectile becomes embedded in the block, the block swings up and back to the position indicated by the dotted rectangle in fig. (6.17). Since measuring the height by which the block rises vertically or the angle through which it swings would be less practicable and accurate, the horizontal distance that the block moves is measured by means of a vertical rod (the red line in Fig. (6.17)) attached to the bottom of the block and extending into a sand pit below: as the block swings back, the rod etches a line in the sand.

    Now that we've managed to get through all that, determine the velocity v_0 of the projectile in terms of m, M, ℓ, and x.

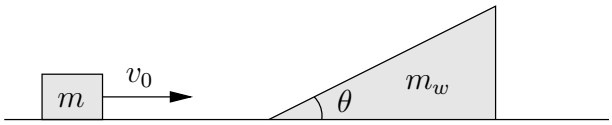


Figure 6.18: Problem 31

31. As shown in fig. (6.18), a small block of mass m, sliding along a frictionless horizontal surface at speed v_0, strikes a stationary wedge of mass m_w and inclination θ. The wedge is free to move along the same frictionless surface, and there is also no friction between the block and the wedge.

    Pretty lame, huh?

    Anyway, you will be concerned with three points in the motion:

    - The initial state, when the block is headed toward the wedge. (This is the state that is illustrated in fig. (6.18).)
    - When the block has reached its highest point on the wedge (which we will assume happens *before* the block gets to the high end of the wedge).
    - The final state, when the block has again separated from the wedge.

    Note that none of the parts of this problem requires much calculation. Think before you calculate.

    (a) What is the initial total momentum of this system (block and wedge)?

    (b) What is the initial total kinetic energy of this system?

    (c) What external forces, if any, act on the system? Is there any net external force on the system?

    (d) Describe what happens to this total momentum and total kinetic energy as the block climbs up the wedge, reaches its highest point, and then descends and separates from the wedge.

    (e) What is the velocity of the center of mass
        i. In the initial state?
        ii. When the block has reached its highest point on the wedge?
        iii. In the final state?

    (f) What is the velocity of the block *relative to the wedge*
        i. In the initial state?
        ii. When the block has reached its highest point on the wedge?
        iii. In the final state?

    (g) How high (measured vertically, from the surface beneath the wedge) is the highest point reached by the block?

    (h) Determine the velocities of the wedge and the block after they again separate.

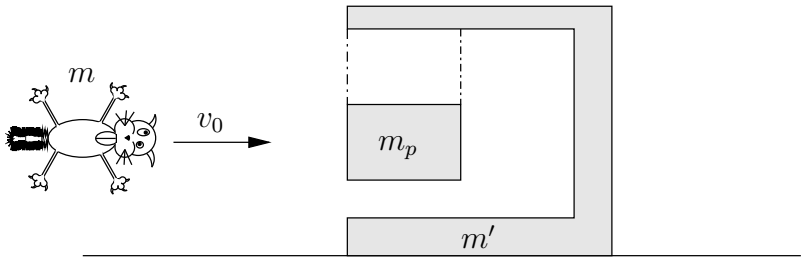


Figure 6.19: Problem 32

32. Fig. (6.19) shows a novel sort of apparatus designed to provide entertainment for scientists: a cat of mass m, fired out of a cannon, strikes a ballistic pendulum with a horizontal velocity v_0 and sticks in the pendulum bob. The dash-dot lines represent the cords by which the pendulum bob (mass m_p) is suspended. The mass of the rest of the apparatus (the framework that supports the bob) is m'. *The apparatus is free to move along a frictionless, level surface.*

    You will be concerned with two points in the motion:

    - The initial state, when the cat is headed toward the pendulum. (This is the state that is illustrated in fig. (6.19).)
    - When the pendulum, with the cat smashed into it, has reached its highest point (which we will assume happens before the pendulum has swung through 90°).

    Also, by *system*, we mean all three bodies together (cat, pendulum bob, and framework).

    (a) What is the initial total momentum of the system?

    (b) What is the initial total kinetic energy of the system?

    (c) Describe what happens to this total momentum and total kinetic energy between the time that the pendulum is struck and the time that the pendulum reaches its highest point.

    (d) What external forces, if any, act on the system? Is there any net external force on the system?

    (e) What is the velocity of the center of mass

        i. In the initial state?

        ii. When the pendulum has reached its highest point?

    (f) How high does the pendulum bob rise (measured vertically, from its original level)?

Continued on next page ...

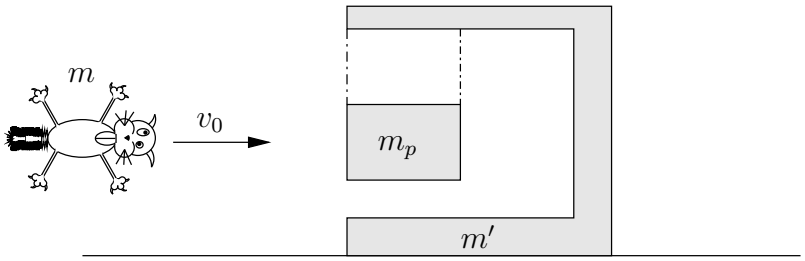


Figure 6.19: Problem 32

(g) How fast is the framework moving when the bob has reached its highest point?

(h) Further into the future, the pendulum bob will, as you would expect, swing back and forth.

i. Qualitatively describe the corresponding motion of the framework as the pendulum swings back and forth.

ii. Will the bob swing as high on the "back" side of its swings as it does on the "front" side?

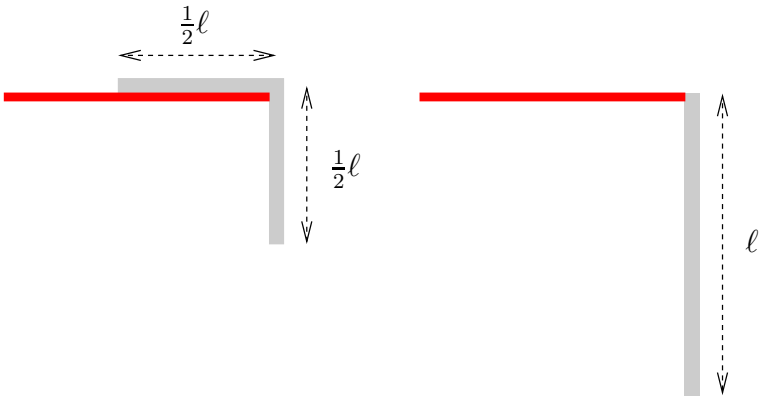


Figure 6.20: Problem 33

33. Fig. (6.20) shows a gray rope sliding off of a red tabletop under the influence of gravity. The tabletop is level and frictionless, and the mass of the rope, which is infinitely flexible, is evenly spread over its length ℓ. Initially the rope is at rest, hanging half on and half off the tabletop, as shown on the left side of fig. (6.20). For simplicity we will assume that the rope stays in a sharp L shape as it slides off the tabletop, so that the part over the edge is always hanging straight vertically downward and the rope is as shown on the right side of fig. (6.20) at the moment it leaves the table.

(a) Determine the speed at which the rope is moving at the moment that it leaves the tabletop (the moment shown on the right side of fig. (6.20)). See the footnote if you need a hint.[19]

(b) Show that, more generally, the speed at which the rope is moving as a function of the length x of rope that is hanging over the side is

$$v = \frac{1}{2}\sqrt{g\ell\left(\left(\frac{2x}{\ell}\right)^2 - 1\right)} \tag{6.28}$$

(c) Set up $F = ma$ for the rope length x of rope hanging over the side of the table and show that the function $x(t)$ of time t that satisfies this relation is

$$x = A\exp\left(\sqrt{\frac{g}{\ell}}\,t\right) + B\exp\left(-\sqrt{\frac{g}{\ell}}\,t\right)$$

where A and B are constants (as yet of undetermined value).

(d) Show that when the conditions that the rope is initially at rest and initially hanging half off the tabletop are imposed on this solution for x, you obtain $A = B = \frac{1}{4}\ell$, so that

$$x = \tfrac{1}{4}\ell\left(\exp\left(\sqrt{\frac{g}{\ell}}\,t\right) + \exp\left(-\sqrt{\frac{g}{\ell}}\,t\right)\right) = \tfrac{1}{2}\ell\cosh\sqrt{\frac{g}{\ell}}\,t \tag{6.29}$$

(e) Show that the velocity yielded by the $x(t)$ of eq. (6.29) satisfies eq. (6.28). It may help to remember that

$$\frac{d}{d\alpha}\cosh\alpha = \sinh\alpha \qquad \frac{d}{d\alpha}\sinh\alpha = \cosh\alpha \qquad \cosh^2\alpha - \sinh^2\alpha = 1$$

(f) The assumption that the rope will stay in a sharp L shape is of course bogus. What really happens to the shape of the rope as it slides off the tabletop?

34. Your roommate settles an argument by squirting you in the face with a fire extinguisher. Determine the force that the stream of water from the extinguisher exerts on your face if it has density ρ and cross-sectional area A, strikes your face at speed v_0, and comes essentially to rest after it hits you. See the footnote if you need a hint.[20]

[19] The gravitational potential energy depends on the location of the center of mass.
[20] Use $F = dp/dt$ and think in terms of time rates.

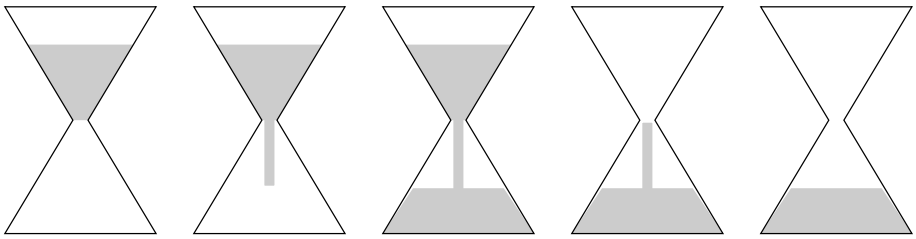


Figure 6.21: Problem 35

35. Fig. (6.21) shows five stages during the draining of sand in an hourglass:

    I. The glass has just been turned over and sand has not yet begun to fall.

    II. Sand has begun to fall, but the column of falling sand has not yet hit the bottom.

    III. Sand is falling and the column of falling sand has reached the bottom.

    IV. The last of the sand is airborne.

    V. All of the sand has fallen and lies in the bottom of the hourglass.

    The hourglass has total mass m (including the sand) and is sitting on a table that, as tables are wont to do, supports it with a normal force. The rate at which the sand falls in mass per unit time is μ (that is, $\mu = dm_{sand}/dt$, where dm_{sand} is the mass of sand that falls in time dt), and we will assume that the sand falls from rest from the neck of the hourglass.

    Determine the normal force exerted on the hourglass by the table

    (a) During stage I.

    (b) During stage II, taking the height of the column of falling sand at the instant in question to be ℓ. See the footnote if you need a hint.[21]

    (c) During stage III. Although you should find that your final result for the normal force does not depend on it, you may assume that the height of the airborne column of sand at the instant in question is ℓ. And don't forget to take into account the impact of the sand that is reaching the bottom.

    (d) During stage IV, taking the height of the column of falling sand at the instant in question to be ℓ and the distance from the neck of the hourglass to the impact point to be h.

    (e) During stage V.

[21] The time rate at which mass is falling is given, and you can figure out how long it took the sand at the lower end of the column to fall through the distance ℓ.

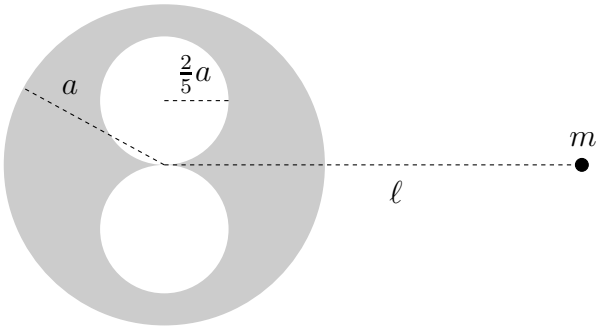


Figure 6.22: Problem 36

36. A point mass m loiters suspiciously a distance ℓ from the center of a sphere of radius a that is composed of a substance of uniform density (mass per unit volume) ρ. From this sphere have been cut out two spherical holes, each of radius $\frac{2}{5}a$, as shown in fig. (6.22). Determine the net gravitational force exerted on the point mass.

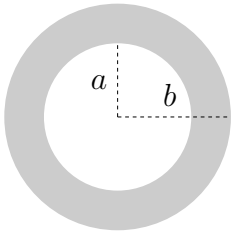


Figure 6.23: Problem 37

37. The hollow sphere shown in fig. (6.23) is composed of a substance of uniform density and has inner radius a, outer radius b, and mass m.

    (a) Determine the gravitational potential energy, measured from out at infinity, of a point mass m' at distances

        i. $r \geq b$ from the center of the sphere.

        ii. $a \leq r \leq b$ from the center of the sphere.

        iii. $r \leq a$ from the center of the sphere.

    (b) How would your results for #37a change if, instead of measuring your gravitational potential energies from out at infinity, you measured them from the center of the sphere?

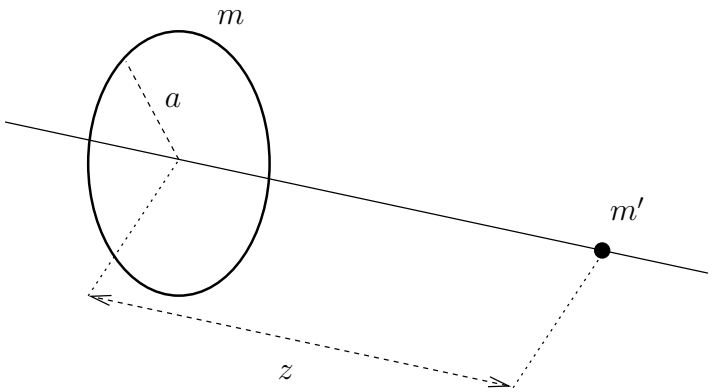


Figure 6.24: Problem 38

38. Fig. (6.24) shows a circular ring of radius a, around which a mass m is evenly distributed. A point mass m' lies on the axis of the ring, a distance z from its center.

    (a) Determine the gravitational potential energy of the point mass due to the ring. Note that it is not necessary to do an integration.

    (b) Determine the net gravitational force exerted on the point mass by the ring. See the footnote if you need a hint.

    (c) Make physical sense of the behavior of your results for the force and potential energy in the limits

        i. $a \to 0$.

        ii. z is large.

    (d) Is it valid to obtain a result for the net gravitational force by using your result of #38a for the gravitational potential energy in eq. (5.13),

    $$F = -\frac{dU}{dz} \quad ?$$

    If so, show that this yields the same net gravitational force that you got in #38b. If not, explain why not.

    (e) Is it valid, working three-dimensionally, to obtain a vector result for the net gravitational force by using your result of #38a for the gravitational potential energy in eq. (5.14),

    $$F = -\nabla U \quad ?$$

    If so, show that this yields the same net gravitational force that you got in #38b. If not, explain why not.

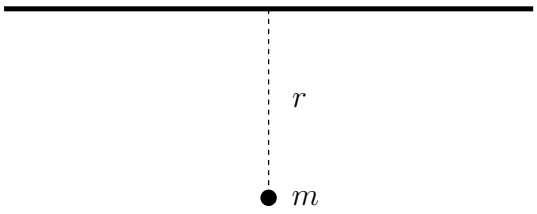


Figure 6.25: Problem 39

39. Fig. (6.25) is supposed to show a point mass m a perpendicular distance r from an infinite rod of uniform linear density λ, but it's difficult to draw an infinite rod, so you'll just have to imagine that the ends of the rod extend out to infinity. Anyway, your task is to determine the net gravitational force exerted on the point mass by the rod. You might find it helpful to remember that, by a tangent substitution,

$$\int \frac{dx}{(a^2+x^2)^{\frac{3}{2}}} = \frac{1}{a}\arctan\frac{x}{a}$$

6.11 Sketchy Answers

(1a) The 45-cm mark.

(2) What you get will depend on where you are measuring from, but the distance $\frac{1}{62}\ell$ should be involved in one way or another.

(3) The distance from the center is $\frac{1}{36}a$.

(4a) What you get will depend on where you are measuring from, but the distance $\dfrac{\ell\sqrt{\ell^2-\frac{1}{4}a^2}}{2\ell+a} = \frac{1}{2}\ell\sqrt{\dfrac{\ell-\frac{1}{2}a}{\ell+\frac{1}{2}a}}$ should figure prominently in your answer.

(5) If you measure from the bottom, $\dfrac{\ell^2}{4(\ell+a)}$.

(6b) $\frac{1}{2}\alpha\ell^2$.

(6c) $x_{cm} = \frac{2}{3}\ell$.

(7a) $\dfrac{2m}{\alpha\ell^2}$.

(7b) $\frac{1}{3}\ell\,(2\hat{\mathbf{x}} + \alpha\hat{\mathbf{y}})$.

(8a) $\dfrac{2m}{\pi a}$.

(8b) $\dfrac{2}{\pi}\,a$.

(8c) If you are looking here, you are thinking about this the wrong way.

(9a) $\dfrac{4m}{\pi(b^2-a^2)}$.

(9b) $\dfrac{4(b^3-a^3)}{3\pi(b^2-a^2)} = \dfrac{4(b^2+ab+a^2)}{3\pi(b+a)}$.

(10a) $\dfrac{2m}{\pi a^2}$.

(10b) $x_{cm} = y_{cm} = z_{cm} = \frac{1}{2}a$.

(11a) $\dfrac{m}{m+m'}\ell$.

(12) With the obvious choice of coordinate axes, $\dfrac{ma}{7m+m'}(\hat{\mathbf{x}} - 3\hat{\mathbf{y}})$.

(15a) 0.085 m/s.

(15b) 0.085 m/s.

(16(a)i) Now that you've worked out an answer, you can check it against the result for the general case.

(16(b)i) Speed $\frac{1}{2}v$ and kinetic energy lost $\frac{5}{4}mv^2$.

(17(a)i) $\frac{5}{6}v$.

(17(a)ii) $\frac{5}{12}mv^2$.

(17b) Now that you've worked out an answer, you can check it against the result for the general case.

(17c) When they stick together: speed $\frac{1}{6}v$, kinetic energy lost $\frac{5}{12}mv^2$. You can check the elastic case against the general elastic result.

(18) $\dfrac{9mV^2}{8\ell}$.

(25a) $v\sqrt{\dfrac{m(m+m')}{km'}}$.

(27a) $\frac{2\sqrt{2}}{3}v$.

(30) Just because the answer is butt-ugly doesn't mean you did something wrong:

$$\frac{m+M}{m}\sqrt{2g\ell\Big(1-\sqrt{1-(x/\ell)^2}\Big)}$$

(32f) $\dfrac{m^2m'}{(m'+m+m_p)(m+m_p)^2}\,\dfrac{v_0^2}{2g}$.

(33a) $\frac{1}{2}\sqrt{3g\ell}$.

(34) $\rho A v_0^2$.

(36) $\dfrac{\frac{4}{3}\pi G\rho m a^3}{\ell^2}\left(1-2\Big(\frac{2}{5}\Big)^3\,\dfrac{\ell^3}{\Big(\ell^2+\Big(\frac{2}{5}a\Big)^2\Big)^{\frac{3}{2}}}\right)$.

(37(a)ii) $-Gmm'\Big(\dfrac{1}{b}+\dfrac{1}{b^3-a^3}\Big(\frac{1}{2}(b^2-r^2)-a^3\Big(\dfrac{1}{r}-\dfrac{1}{b}\Big)\Big)\Big)$.

(37(a)iii) $-Gmm'\Big(\dfrac{1}{b}+\dfrac{1}{b^3-a^3}\Big(\frac{1}{2}(b^2-a^2)-a^3\Big(\dfrac{1}{a}-\dfrac{1}{b}\Big)\Big)\Big)$.

(38a) $-\dfrac{Gmm'}{\sqrt{a^2+z^2}}$.

(38b) $\dfrac{Gmm'z}{(a^2+z^2)^{\frac{3}{2}}}$.

(39) $\dfrac{2G\lambda m}{r}$.

Chapter 7

Rotational Dynamics

> Whoever said the hand is quicker than the eye
> obviously never tried rolling them down a ramp.
>
> Paul Paternoster

7.1 Two-Dimensional Rotations

Up to now, we have dealt only with translational motion, that is, with the movement of a body from one location to another.[1] Rotational kinematics and dynamics deal with bodies rotating about axes.

It turns out that rotational motion is exactly analogous to translational motion: through a simple set of correspondences, we can rewrite each of the relations governing translational motion in a form that applies to rotational motion, and the reasoning and methods for working with these rotational relations are almost identical to those already familiar to you from translations.

Our first task will be to discover the rotational analogues of the various translational quantities by considering some generic body rotating about an axis. Fig. (7.1) shows a rectangle rotating at angular velocity ω about a perpendicular axis through its center, as seen from above (so that we are looking down the axis of rotation, which is represented by the red dot). This rectangle constitutes what is known as an extended body, which means simply that its mass is distributed over a finite region — as opposed to being concentrated at a point like a point mass. We can regard extended bodies as a collection of point masses $\{m_i\}$, $i = 1, 2, \ldots, n$.[2] All of the point masses

[1] Even our study of circular motion qualifies as translational motion the way we approached it.

[2] For a continuous body like our rectangle, we could and strictly should deal, not with a set of discrete mass points, but with a continuum of infinitesimal bits of mass dm. We use

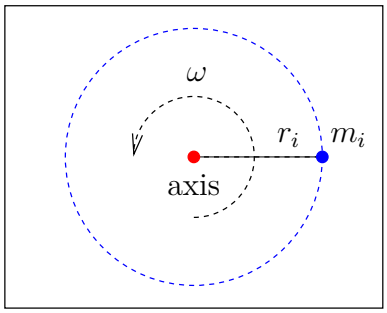


Figure 7.1: A Rotating Rectangle. Whoopee.

will be rotating at the same angular rate ω, but each point mass m_i has its own radius r_i of revolution.

Since each m_i is undergoing the kind of simple circular motion we have already studied, we immediately have several results. The arc length s_i traveled by m_i will be related to the angle θ through which the body has rotated by

$$s_i = \theta r_i \tag{7.1a}$$

Likewise the velocity v_i and tangential acceleration $a_{tan,i}$ of m_i are related to the angular velocity ω and angular acceleration α of the body by

$$v_i = \omega r_i \tag{7.1b}$$

$$a_{tan,i} = \alpha r_i \tag{7.1c}$$

where, as before (in eqq. (3.28) on p.139), angular velocity and acceleration are defined by

$$\omega = \frac{d\theta}{dt} \qquad\qquad \alpha = \frac{d\omega}{dt}$$

As you may recall, eqq. (7.1b) and (7.1c) can in fact be obtained simply by taking time derivatives of eq. (7.1a).

The $\pm$ signs on θ, ω, and α, just as was the case in one-dimensional translational motion, tell us the directions of these quantities — that is, whether they are clockwise or counterclockwise. The convention is to take counterclockwise to be positive and clockwise negative, but you can of course choose to do the reverse; you just have to be clear about which choice you are making and then to be consistent about your signs once you have made that choice.[3]

a set of discrete mass points only because it makes the derivations somewhat simpler; all of the proofs and arguments we will make apply to both discrete and continuous distributions of mass.

[3]In the more general treatment we will be giving in §7.2, we will be making θ, ω, and

Kinetic Energy & Moment of Inertia

Consider now the kinetic energy K of the rotating body. To get the total kinetic energy of the body, we should sum up the kinetic energies of all of the point masses that make up the body:

$$K = \sum_{i=1}^{n} \tfrac{1}{2} m_i v_i^2$$

While this is a perfectly fine and correct expression for the total kinetic energy of the body, the velocities v_i of course differ from one mass point to another; it would be more natural to express the kinetic energy of the body in terms of quantities related to the body and its overall rotational motion rather than to the motion of individual points within the body. And there is a simple way to accomplish this: while the velocity v_i differs from one mass point to another, the angular velocity ω is common to all points in the body, so that, using $v_i = \omega r_i$ from eq. (7.1b), we have

$$\begin{aligned} K &= \sum_{i=1}^{n} \tfrac{1}{2} m_i (\omega r_i)^2 \\ &= \tfrac{1}{2} \left(\sum_{i=1}^{n} m_i r_i^2 \right) \omega^2 \end{aligned}$$

We can write this in the simple form

$$K = \tfrac{1}{2} I \omega^2 \tag{7.2}$$

if we define a new quantity I, known as the moment of inertia, to be

$$I = \sum_{i=1}^{n} m_i r_i^2 \tag{7.3}$$

Comparing eq. (7.2) to the translational relation $K = \frac{1}{2} m v^2$, we see that, since the angular velocity ω is the rotational equivalent of velocity v, this new quantity I is the rotational equivalent of mass. If we regard the body as a continuous distribution of mass rather than a collection of discrete point

α vector quantities: although it is always possible to describe the rotation of a body at a given moment as simply clockwise or counterclockwise, the orientation of the body's axis of rotation may itself be changing orientation as time passes, and θ, ω, and α may therefore differ from each other in direction. The convention is to take the vector for a rotational quantity to be along its axis in a right-handed sense. Thus the direction of the counterclockwise ω in fig. (7.1) would as a vector be out of the page ($\odot$). For the time being, however, we will restrict ourselves to rotational motions for which θ, ω, and α are all along the same axis and can therefore be described simply as clockwise or counterclockwise.

masses, then we have an integration rather than a sum:[4]

$$I = \int_{\text{body}} dm\, r^2 \tag{7.4}$$

In §7.5 we will carry out this integration for bodies of some common symmetric shapes (rings, disks, spheres, etc.).

Torque

Next we tackle the rotational equivalent of force, which is known as torque and denoted by τ. We would like to find a definition for τ that will give us the rotational equivalent of $F = ma$. Since the rotational equivalents of mass and acceleration are the moment of inertia I and the angular acceleration α, this means that we would like to find a definition for τ that will give us

$$\tau = I\alpha \tag{7.5}$$

Using eq. (7.3), and eq. (7.1c) in the form $\alpha = a_{tan,i}/r_i$, in eq. (7.5), we have

$$\tau = \sum_{i=1}^{n} m_i r_i^2 \frac{a_{tan,i}}{r_i} = \sum_{i=1}^{n} r_i m_i a_{tan,i} = \sum_{i=1}^{n} r_i F_{tan,i} \tag{7.6}$$

where we have noted that $m_i a_{tan,i}$ is, according to $F = ma$, the tangential component of the net force on m_i. As you can see from the upper drawing in fig. (7.2), $F_{tan,i} = F_i \sin\phi_i$, so that we have

$$\tau = \sum_{i=1}^{n} r_i F_i \sin\phi_i \tag{7.7}$$

[4] If this is the first time that you have seen the differential written to the left of the integrand, you should get used to it; writing the differential to the right of the integrand is little-league stuff. Sometimes one has to deal with integrations many lines or pages long, and in such cases it would be absurd to postpone the differential, which tells you what variable you are integrating over, until the very end. More importantly, writing the differential next to the integral sign also makes it unambiguously clear which limits go with which integrand in higher-dimensional integrations — as opposed to trying to recall some woefully artificial convention about what goes with what when the differentials are written to the right, which is as hopeless as trying to figure out which fork to use at a formal dinner: in

$$\int_a^b \int_c^d (blah\ blah)\, dx\, dy$$

does x go from a to b and y from c to d, or the other way around? So much better to write

$$\int_a^b dx \int_c^d dy\ (blah\ blah)$$

and then everything is crystal-clear.

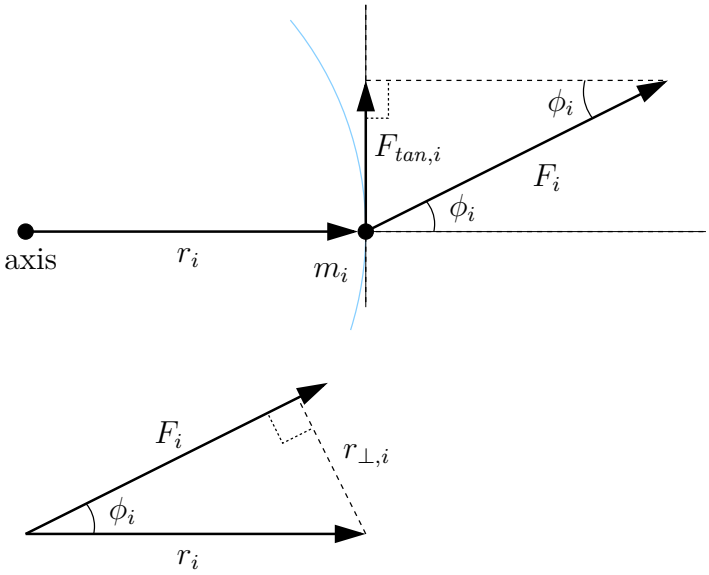


Figure 7.2: $F_{tan,i} = F_i \sin\phi_i$ and $r_{\perp,i} = r_i \sin\phi_i$

The combination $r_i F_i \sin\phi_i$ is just the magnitude of the geometric definition of cross product (see eq. (1.7) on p.74). Thus we arrive at[5]

$$\boldsymbol{\tau} = \sum_{i=1}^{n} \mathbf{r}_i \times \mathbf{F}_i$$

That is, we will indeed have relation (7.5) with τ being the sum over the torques on the point masses that make up the body if we define the torque τ due to a force $\mathbf{F}$ to be

$$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F} \tag{7.8}$$

where the radius $\mathbf{r}$ goes from the axis of rotation to the point where the force $\mathbf{F}$ is acting.

Now that we have a result for torque, some observations:

- $\tau = I\alpha$ is the rotational analogue of $F = ma$. Just as $F = ma$ tells us that the net force on a body determines its translational motion in the sense of determining its linear acceleration a, $\tau = I\alpha$ tells us that the net torque on a body determines its rotational motion in the sense of determining its angular acceleration α. The rotational analogue of

[5]Yes, we have pulled a bit of a fast one here by suddenly making τ into a vector. But no harm is done: for the simple cases where the rotational motion is either clockwise or counterclockwise in the plane of the page, the corresponding vector torques in the above cross product will be, by the right-hand rule, either into or out of the page, respectively. The vector cross product will therefore make clockwise and counterclockwise opposite in direction, entirely equivalent to accounting for these directions by $\pm$ signs.

the inertial mass m that resists changes in translational motion is the moment of inertia I, which resists changes in rotational motion.

- Since $I = \int dm\, r^2$, a body's moment of inertia and resistance to changes in rotational motion depend not only on how massive it is, but on how that mass is distributed: the farther it is from the axis of rotation (that is, the larger its radius r), the greater the contribution the mass makes to the moment of inertia. Thus a hoop will have a larger moment of inertia and be harder to set rotating than a solid disk of the same mass, because the hoop's mass is farther from its center.

- $\tau = rF_{tan}$ (from eq. (7.6)) tells us that only the tangential component of a force contributes to the torque. This should seem perfectly plausible: to the extent that a force is radial, it is just pushing or pulling directly toward or away from the axis of rotation and is not contributing at all to any rotational motion around that axis.

- $\tau = rF_{tan}$ also tells us that the torque generated by a force is proportional to the distance r from the axis of rotation to the point where the force is applied. This is why you apply force to the outer end of wrench when you're trying to turn a bolt.

Although $\tau = rF_{tan}$ is helpful for understanding what torque corresponds to and means physically, there is another expression for τ that is usually more convenient for calculations. Backing up to eq. (7.7), you can see from the lower drawing in fig. (7.2) that the combination $r_i F_i \sin\phi_i$ can equally well be expressed as $r_{\perp,i} F_i$. Thus we can also express the τ as

$$\tau = r_\perp F \tag{7.9}$$

The $r_\perp$, called the torque arm, is the perpendicular distance to the line along which the force F is acting — what we might informally call the "line of force," although this isn't a technical term. We'll postpone a concrete example until we get to §7.10, in which we'll work out the case of rolling in detail, but the general procedure for evaluating a torque by means of eq. (7.9) is to draw the line of force by extending the force vector, then drop a perpendicular from the axis of rotation to this line. This perpendicular segment is $r_\perp$, as shown in fig. (7.3).

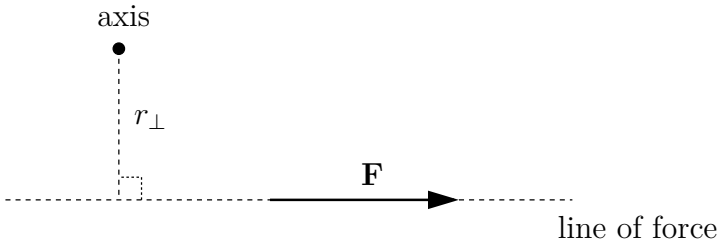


Figure 7.3: Using $\tau = r_\perp F$

Angular Momentum

Finally, we need to determine the rotational equivalent of the linear momentum $p = mv$. Since the rotational equivalents of mass m and linear velocity v are the moment of inertia I and angular velocity ω, we would like to find a definition of angular momentum L for which

$$L = I\omega$$

We could go through the same chain of reasoning that we did for torque, but there is a lazier way to get what we need: for torque, we set out to find a definition of τ that would yield $\tau = I\alpha$ and ended up with

$$\tau = |\mathbf{r} \times \mathbf{F}| = r_\perp F \tag{7.10}$$

Now, the angular velocity ω in $L = I\omega$ is the antiderivative[6] of the angular acceleration α in $\tau = I\alpha$. So we expect that the angular momentum L will be the antiderivative of the torque τ and that therefore we will have for L relations like (7.10), but involving the antiderivative of F in place of the force F. If we recall that

$$F = \frac{dp}{dt}$$

we see that momentum is the antiderivative of force and that we should therefore have

$$\mathbf{L} = \mathbf{r} \times \mathbf{p} \tag{7.11a}$$

$$L = |\mathbf{r} \times \mathbf{p}| = r_\perp p \tag{7.11b}$$

These are in fact the defining relations for angular momentum.

Note that, according to $L = I\omega$,

$$\frac{dL}{dt} = \frac{d(I\omega)}{dt} = I\frac{d\omega}{dt} = I\alpha = \tau$$

[6] Antiderivative is another cool word. It sounds so demonic.

so that, just as we expected, torque is indeed the rate of change of angular momentum,

$$\tau = \frac{dL}{dt} \tag{7.12}$$

Recall that when there was no net external force on a system, its total linear momentum was conserved. The rotational analogue of this is that when there is no net external torque on a system, its angular momentum is conserved. We'll say more about conservation of angular momentum in §7.7.

Work & Power

To get a result for the work done by a torque, we start from the definition of the work dW done by a force $\mathbf{F}$ acting through an infinitesimal displacement $d\mathbf{r}$,

$$dW = \mathbf{F} \cdot d\mathbf{r}$$

and note that, as always, work is done by the force $\mathbf{F}$ only to the extent that the force is along (that is, parallel or antiparallel) to the displacement $d\mathbf{r}$. Since the displacement of a rotating body is in the tangential direction, only the tangential component of $\mathbf{F}$ does work. The magnitude of the displacement $d\mathbf{r}$ being simply the arc length $ds = r\,d\theta$, we thus have

$$\begin{aligned} dW &= F_{tan}\,ds \\ &= F_{tan} r\,d\theta \\ &= \tau\,d\theta \end{aligned} \tag{7.13}$$

where in the last step we have used $\tau = rF_{tan}$ from eq. (7.6). The work W done by a torque τ is therefore

$$W = \int \tau\,d\theta$$

For a constant torque, this reduces to

$$W = \tau\theta \tag{7.14}$$

where θ is the angle through which the torque has acted. Eq. (7.14) is the rotational analog of $W = Fs$ for a constant one-dimensional force F.

To obtain a result for the power associated with a torque, we just use eq. (7.13) and $\omega = d\theta/dt$ in conjunction with the basic definition of power as the rate of doing work:

$$P = \frac{dW}{dt}$$

$$= \tau \frac{d\theta}{dt}$$
$$= \tau\omega \tag{7.15}$$

Eq. (7.15) is the rotational analog of $P = Fv$ for one-dimensional translational motion.

Summary

Table (7.1) summarizes the correspondences between translational and rotational kinematical and dynamical quantities. The various rotational relations are applied with exactly the same reasoning and methods already familiar to you from translational motion — almost. The one difference is that bodies may no longer be represented simply by points in force diagrams. Up to now in the diagrams we have been drawing for the purpose of applying $F = ma$ to the translational motion of bodies, it did *not* matter where on the body a force acted, and we were therefore able to reduce the body to a point for simplicity. But since the value of the torque *does* depend on where the force giving rise to that torque is applied, when drawing force diagrams for applying $\tau = I\alpha$ to the rotational motion of bodies we must be careful to draw the tail of each force vector from the point where that force actually (or effectively) acts on the body. We'll postpone an example of this until we get to the case of rolling in §7.10, which we'll work out in detail.

Translational		Rotational		Relation Between
Location	x	Angle of orientation	θ	$s = \theta r$ (s = arc length)
Velocity	v	Angular velocity	ω	$v = \omega r$
Acceleration	a	Angular acceleration	α	$a = \alpha r$
Mass	m	Moment of inertia	I	$I = \sum m_i r_i^2$ or $I = \int dm\, r^2$
Force	$F = ma$	Torque	$\tau = I\alpha$	$\tau = r_\perp F = rF_{tan}$
	$= \frac{dp}{dt}$		$= \frac{dL}{dt}$	
Momentum	$p = mv$	Angular momentum	$L = I\omega$	$L = r_\perp p$
Kinetic energy	$K = \frac{1}{2}mv^2$	Kinetic energy	$K = \frac{1}{2}I\omega^2$	
Work	$W = Fs$	Work	$W = \tau\theta$	
Power	$P = Fv$	Power	$P = \tau\omega$	

Table 7.1: Rotational Equivalents of Translational Quantities

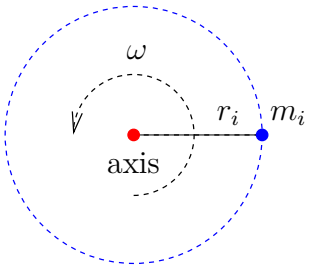


Figure 7.4: Déjà Vu All Over Again

7.2 Three-Dimensional Rotations

In this section we will be repeating the derivations of the various rotational quantities and the relations among them for rotations that cannot be reduced to simply clockwise and counterclockwise. In this more general case, the angle $\boldsymbol{\theta}$, angular velocity $\boldsymbol{\omega}$, and angular acceleration $\boldsymbol{\alpha}$ are all vector quantities, with the vector for each pointing along its rotational axis in a right-handed sense. Thus the direction of the counterclockwise ω in fig. (7.4), for example, is out of the page ($\odot$).

We begin by noting that the rotational velocity $\mathbf{v}_i$ of a mass point m_i is related to its angular velocity $\boldsymbol{\omega}$ by

$$\mathbf{v}_i = \boldsymbol{\omega} \times \mathbf{r}_i \tag{7.16}$$

This works for the m_i shown in fig. (7.4): $\boldsymbol{\omega}$ is $\odot$ and $\mathbf{r}_i$ is $\rightarrow$, so that the direction of $\boldsymbol{\omega} \times \mathbf{r}_i$ is indeed $\uparrow$, and since there is a right angle between $\boldsymbol{\omega}$ and $\mathbf{r}_i$ the magnitudes also check out:

$$v_i = \omega r_i \sin\frac{\pi}{2} = \omega r_i$$

Fig. (7.5) shows, from a somewhat medieval perspective, that eq. (7.16) also holds in more general cases:[7] m_i is rotating around the axis with a velocity $\mathbf{v}_i$ that is, at the moment shown in the figure, into the page ($\otimes$), which is again the same as the direction of $\boldsymbol{\omega} \times \mathbf{r}_i$. And the magnitudes also still check out:

$$v_i = \omega r_i \sin\varphi = \omega r_{\perp,i}$$

[7] We are still restricting ourselves to coordinate systems that have their origins somewhere on the axis of rotation, but, since this makes our lives simpler without loss of generality, we will always be using such coordinate systems when we do rotational motion.

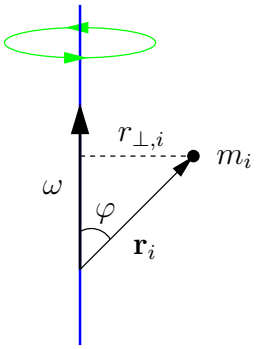


Figure 7.5: $\mathbf{v}_i = \boldsymbol{\omega} \times \mathbf{r}_i$

Kinetic Energy & Moment of Inertia

From eq. (7.16) we can derive a more general result than eq. (7.2) for the kinetic energy of a rotating body. Starting from

$$K = \sum_{i=1}^{n} \tfrac{1}{2} m_i v_i^2 = \sum_{i=1}^{n} \tfrac{1}{2} m_i (\boldsymbol{\omega} \times \mathbf{r}_i)^2$$

and using the general vector relation [8]

$$(\mathbf{a} \times \mathbf{b}) \cdot (\mathbf{c} \times \mathbf{d}) = \mathbf{a} \cdot \mathbf{c}\, \mathbf{b} \cdot \mathbf{d} - \mathbf{a} \cdot \mathbf{d}\, \mathbf{b} \cdot \mathbf{c}$$

we have

$$\begin{aligned} K &= \sum_{i=1}^{n} \tfrac{1}{2} m_i \Big(\boldsymbol{\omega} \cdot \boldsymbol{\omega}\, \mathbf{r}_i \cdot \mathbf{r}_i - (\boldsymbol{\omega} \cdot \mathbf{r}_i)^2 \Big) \\ &= \sum_{i=1}^{n} \tfrac{1}{2} m_i \Big(\omega^2 r_i^2 - (\boldsymbol{\omega} \cdot \mathbf{r}_i)^2 \Big) \end{aligned}$$

We would like to pull the ω, which does not depend on the index i, outside of the sum, but this is not as easy as just factoring the ω out because in the second term it is involved in a dot product. To pull the ω out, we need to work with matrices: if you've dealt with matrices in some previous life, you can see that the above expression for K can be written as

$$\tfrac{1}{2}\, \boldsymbol{\omega} \cdot I \cdot \boldsymbol{\omega} \tag{7.17}$$

if we define I to be the matrix

$$I = \sum_{i=1}^{n} m_i \begin{pmatrix} r_i^2 - x_i^2 & -x_i y_i & -x_i z_i \\ -x_i y_i & r_i^2 - y_i^2 & -y_i z_i \\ -x_i z_i & -y_i z_i & r_i^2 - z_i^2 \end{pmatrix} \tag{7.18}$$

[8] You can verify this relation simply by expanding both sides — it's a pain, but straightforward.

To prove this, we just expand (7.17):

$$\begin{aligned}
\tfrac{1}{2}\,\boldsymbol{\omega}\cdot I\cdot\boldsymbol{\omega} &= \tfrac{1}{2}\begin{pmatrix}\omega_x & \omega_y & \omega_z\end{pmatrix}\cdot \\
&\qquad \sum_{i=1}^{n} m_i \begin{pmatrix} r_i^2-x_i^2 & -x_iy_i & -x_iz_i \\ -x_iy_i & r_i^2-y_i^2 & -y_iz_i \\ -x_iz_i & -y_iz_i & r_i^2-z_i^2 \end{pmatrix}\cdot\begin{pmatrix}\omega_x\\ \omega_y\\ \omega_z\end{pmatrix} \\
&= \tfrac{1}{2}\begin{pmatrix}\omega_x & \omega_y & \omega_z\end{pmatrix}\cdot \\
&\qquad \sum_{i=1}^{n} m_i \begin{pmatrix} \omega_x(r_i^2-x_i^2)-\omega_y x_iy_i-\omega_z x_iz_i \\ -\omega_x x_iy_i+\omega_y(r_i^2-y_i^2)-\omega_z y_iz_i \\ -\omega_x x_iz_i-\omega_y y_iz_i+\omega_z(r_i^2-z_i^2) \end{pmatrix} \\
&= \tfrac{1}{2}\sum_{i=1}^{n} m_i\Bigg(\omega_x\Big(\omega_x(r_i^2-x_i^2)-\omega_y x_iy_i-\omega_z x_iz_i\Big) \\
&\qquad +\omega_y\Big(-\omega_x x_iy_i+\omega_y(r_i^2-y_i^2)-\omega_z y_iz_i\Big) \\
&\qquad +\omega_z\Big(-\omega_x x_iz_i-\omega_y y_iz_i+\omega_z(r_i^2-z_i^2)\Big)\Bigg)
\end{aligned}$$

This is not as bad as it looks: the terms involving r_i^2 combine into

$$\tfrac{1}{2}\sum_{i=1}^{n} m_i(\omega_x^2+\omega_y^2+\omega_z^2)r_i^2 = \sum_{i=1}^{n}\tfrac{1}{2}m_i\omega^2 r_i^2$$

and the other terms into

$$-\tfrac{1}{2}\sum_{i=1}^{n} m_i(\omega_x x_i+\omega_y y_i+\omega_z z_i)^2 = \sum_{i=1}^{n} -\tfrac{1}{2}m_i(\boldsymbol{\omega}\cdot\mathbf{r}_i)^2$$

Thus the matrix multiplication (7.17) with the matrix I of eq. (7.18) does indeed reproduce the K we want.

The I of eq. (7.18) is known as the inertia tensor. Scalars, vectors, etc., are defined by their transformation properties under operations like spatial rotations and parity. Vectors have a single index in the sense that they are column vectors whose components can be denoted by a single subscript (a_x, a_y, and a_z) and thus constitute tensors of rank one. Matrices like I have two indices in the sense that their elements are denoted by two subscripts ($I_{xx} = r_i^2 - x_i^2$, $I_{xy} = -x_iy_i$, ...) and constitute tensors of rank two. A tensor of rank two is in a sense just two vectors smashed together.

Anyway, for continuous distributions of mass, the discrete sum becomes an integral, so that the inertia tensor is given by

$$I = \int dm \begin{pmatrix} r^2-x^2 & -xy & -xz \\ -xy & r^2-y^2 & -yz \\ -xz & -yz & r^2-z^2 \end{pmatrix} \tag{7.19}$$

Now that we've gone to all the trouble of deriving the inertia tensor for three-dimensional rotations, the first thing we will do with it is show that it reduces to eq. (7.4) for two-dimensional rotations. If we take the z axis to be the axis of rotation, then $\boldsymbol{\omega} = \omega\,\widehat{\mathbf{z}}$, so that the expression (7.17) for the kinetic energy reduces to

$$\begin{aligned}
K &= \tfrac{1}{2}\begin{pmatrix}0 & 0 & \omega\end{pmatrix} \cdot \int dm \begin{pmatrix} r^2-x^2 & -xy & -xz \\ -xy & r^2-y^2 & -yz \\ -xz & -yz & r^2-z^2 \end{pmatrix} \cdot \begin{pmatrix}0\\0\\\omega\end{pmatrix} \\
&= \tfrac{1}{2}\begin{pmatrix}0 & 0 & \omega\end{pmatrix} \cdot \int dm\, \omega \begin{pmatrix} -xz \\ -yz \\ r^2-z^2 \end{pmatrix} \\
&= \tfrac{1}{2}\int dm\, \omega^2 (r^2 - z^2) \\
&= \tfrac{1}{2}\int dm\, \omega^2 \Big((x^2+y^2+z^2) - z^2 \Big) \\
&= \tfrac{1}{2}\omega^2 \int dm\, (x^2+y^2) \\
&= \tfrac{1}{2}\omega^2 \int dm\, r_\perp^2
\end{aligned}$$

where

$$r_\perp^2 = x^2 + y^2$$

is the perpendicular distance of the mass element dm from the axis of rotation (the z axis). Thus the inertia tensor reduces, for two-dimensional rotations, to

$$I = \int dm\, r_\perp^2 \tag{7.20}$$

which is indeed the same as eq. (7.4).

Angular Momentum & Torque

We will begin by defining the angular momentum of a mass m to be

$$\mathbf{L} = \mathbf{r} \times \mathbf{p} \tag{7.21}$$

where $\mathbf{p}$ is the usual linear momentum $m\mathbf{v}$. If there are several discrete masses, the total angular momentum will be

$$\mathbf{L} = \sum_{i=1}^{n} \mathbf{r}_i \times \mathbf{p}_i$$

and if there is a continuous distribution of mass it will be

$$\mathbf{L} = \int \mathbf{r}_i \times dm\, \mathbf{v}_i$$

To show that definition (7.21) yields

$$\boldsymbol{\tau} = \frac{d\mathbf{L}}{dt} \tag{7.22}$$

and

$$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F} \tag{7.23}$$

all we have to do is take a time derivative of (7.21):

$$\begin{aligned}
\frac{d\mathbf{L}}{dt} &= \frac{d}{dt}(\mathbf{r} \times \mathbf{p}) \\
&= \frac{d}{dt}(\mathbf{r} \times m\mathbf{v}) \\
&= \frac{d\mathbf{r}}{dt} \times m\mathbf{v} + \mathbf{r} \times m\frac{d\mathbf{v}}{dt} \\
&= \mathbf{v} \times m\mathbf{v} + \mathbf{r} \times m\mathbf{a} \\
&= 0 + \mathbf{r} \times \mathbf{F}
\end{aligned}$$

where we have noted that $\mathbf{v} \times \mathbf{v} = 0$, as is true for any vector crossed into itself. So if we define torque by eq. (7.23) and angular momentum by eq. (7.21), eq. (7.22) follows.

Note that both angular momentum and torque depend on the vector $\mathbf{r}$, which is measured from a point that we may choose arbitrarily.[9] By shifting our choice of the point about which we evaluate our torques and angular momenta, we change the values of these torques and angular momenta. It is therefore not the values of torques and angular momenta per se that have meaning, but rather their adherence to relations like eq. (7.22).

Note also that, according to $\mathbf{L} = \mathbf{r} \times \mathbf{p}$, even objects that are not rotating — such as a mass moving along a straight line — can still have nonzero angular momenta.

The relations (7.21) and (7.22) apply to any kind of motion. We now want to show that for purely rotational motion of a body about an axis, we also have

$$\mathbf{L} = I \cdot \boldsymbol{\omega} \tag{7.24}$$

and

$$\boldsymbol{\tau} = I \cdot \boldsymbol{\alpha} \tag{7.25}$$

with I being the inertia tensor (7.18) or (7.19). For simplicity, we will look only at the case of a single mass point m, but the proofs are virtually identical for collections of discrete mass points and for continuous distributions of mass.

[9] Though of course we must be consistent: once we have chosen a point, we must evaluate all angular momenta and torques about this same point.

For a single mass point m, eqq. (7.16) and (7.21) give

$$\mathbf{L} = \mathbf{r} \times \mathbf{p} = \mathbf{r} \times m\mathbf{v} = \mathbf{r} \times m(\boldsymbol{\omega} \times \mathbf{r})$$

If we use the general vector relation

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{a} \cdot \mathbf{c}\, \mathbf{b} - \mathbf{a} \cdot \mathbf{b}\, \mathbf{c}$$

then this becomes

$$\mathbf{L} = m(\mathbf{r} \cdot \mathbf{r}\, \boldsymbol{\omega} - \mathbf{r} \cdot \boldsymbol{\omega}\, \mathbf{r}) = m(r^2\, \boldsymbol{\omega} - \mathbf{r} \cdot \boldsymbol{\omega}\, \mathbf{r})$$

Writing this as a column vector, and using

$$\mathbf{r} \cdot \boldsymbol{\omega} = x\omega_x + y\omega_y + z\omega_z$$

we have

$$\begin{aligned}
\mathbf{L} &= m \begin{pmatrix} r^2\omega_x - \mathbf{r} \cdot \boldsymbol{\omega}\, x \\ r^2\omega_y - \mathbf{r} \cdot \boldsymbol{\omega}\, y \\ r^2\omega_z - \mathbf{r} \cdot \boldsymbol{\omega}\, z \end{pmatrix} \\
&= m \begin{pmatrix} r^2\omega_x - (x\omega_x + y\omega_y + z\omega_z)\, x \\ r^2\omega_y - (x\omega_x + y\omega_y + z\omega_z)\, y \\ r^2\omega_z - (x\omega_x + y\omega_y + z\omega_z)\, z \end{pmatrix} \\
&= m \begin{pmatrix} (r^2 - x^2)\omega_x - xy\omega_y - xz\omega_z \\ -xy\omega_y + (r^2 - y^2)\omega_y - yz\omega_z \\ -xz\omega_z - yz\omega_z + (r^2 - z^2)\omega_z \end{pmatrix} \\
&= m \begin{pmatrix} r^2 - x^2 & -xy & -xz \\ -xy & r^2 - y^2 & -yz \\ -xz & -yz & r^2 - z^2 \end{pmatrix} \cdot \begin{pmatrix} \omega_x & \omega_y & \omega_z \end{pmatrix}
\end{aligned}$$

which is indeed just $I \cdot \boldsymbol{\omega}$.

Though we won't go through all the details, to establish that $\boldsymbol{\tau} = I \cdot \boldsymbol{\alpha}$, we would start from

$$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F} = \mathbf{r} \times m\mathbf{a}$$

and, using $\mathbf{a} = \boldsymbol{\alpha} \times \mathbf{r}$, go through the same sequence of steps that we did above for $\mathbf{L} = I \cdot \boldsymbol{\omega}$.

Work & Power

To get a result for power, we start from

$$P = \mathbf{F} \cdot \mathbf{v}$$

and once again use eq. (7.16):

$$P = \mathbf{F} \cdot \mathbf{v} = \mathbf{F} \cdot (\boldsymbol{\omega} \times \mathbf{r})$$

This combination of dot and cross products is known as a scalar triple product and has the general property that we can freely switch the dot and cross, can freely rearrange the order of the vectors cyclically, and, with a negative sign difference, can rearrange the order of the vectors acyclically: [10]

$$\begin{aligned} \mathbf{a} \times \mathbf{b} \cdot \mathbf{c} = &\quad \mathbf{a} \cdot \mathbf{b} \times \mathbf{c} \\ = &\quad \mathbf{c} \times \mathbf{a} \cdot \mathbf{b} = \quad \mathbf{b} \times \mathbf{c} \cdot \mathbf{a} \\ = &-\mathbf{b} \times \mathbf{a} \cdot \mathbf{c} = -\mathbf{c} \times \mathbf{b} \cdot \mathbf{a} = -\mathbf{a} \times \mathbf{c} \cdot \mathbf{b} \end{aligned}$$

Thus we have

$$P = \mathbf{r} \times \mathbf{F} \cdot \boldsymbol{\omega}$$

or

$$P = \boldsymbol{\tau} \cdot \boldsymbol{\omega} \tag{7.26}$$

Since

$$P = \frac{dW}{dt} \qquad \text{and} \qquad \boldsymbol{\omega} = \frac{d\boldsymbol{\theta}}{dt}$$

eq. (7.26) can be written

$$\frac{dW}{dt} = \boldsymbol{\tau} \cdot \frac{d\boldsymbol{\theta}}{dt}$$

Integrating both sides with respect to time, we obtain for the work done by a torque

$$W = \int \boldsymbol{\tau} \cdot d\boldsymbol{\theta} \tag{7.27}$$

If the torque is constant, this becomes

$$W = \boldsymbol{\tau} \cdot \boldsymbol{\theta} \tag{7.28}$$

where $\boldsymbol{\theta}$ is the angular displacement.

[10] To prove this, you have only to do out the cross and then the dot products — very tedious, but straightforward.

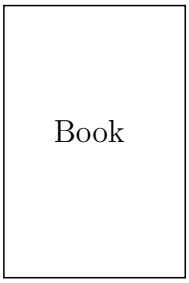


Figure 7.6: What You See is What You Get

Weird Properties of Three-Dimensional Rotations

One complication of three-dimensional rotations is that they are non-Abelian, that is, they do not commute with each other. In two dimensions, where the only issue is clockwise and counterclockwise, rotations do commute, so the order in which you perform a sequence of rotations does not matter. But in three or more dimensions this is no longer true. To see that this is so, try holding a book with the cover facing you, as rather lamely shown in fig. (7.6). If you first rotate the book by $\boldsymbol{\theta}_1 = \frac{\pi}{2} \uparrow$ and then by $\boldsymbol{\theta}_2 = \frac{\pi}{2} \rightarrow$, you should find yourself looking at the top of the book. But if you do these rotations in the opposite order, you should instead find yourself looking at the spine of the book. Freaky.

Another complication of three-dimensional rotations is that quantities like angular momentum and angular velocity need not be parallel to each other. As a simple example of this, consider a point mass m rotating about the z axis, at a constant angular velocity ω, at the end of a massless rod of length ℓ at $\frac{\pi}{4}$ to the z axis, as shown in fig. (7.7). The location of the mass

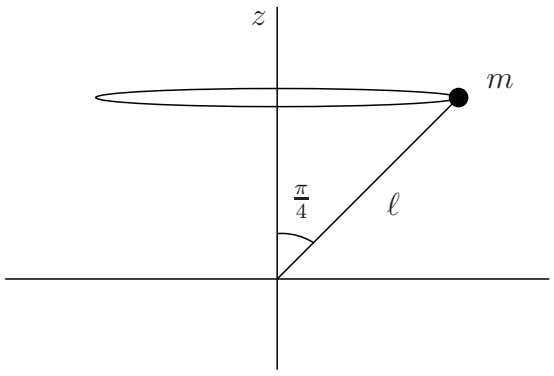


Figure 7.7: Rotational Weirdness

will be given by something like[11]

$$x = r_\perp \cos\omega t = \tfrac{1}{\sqrt{2}}\ell\cos\omega t \qquad y = r_\perp \sin\omega t = \tfrac{1}{\sqrt{2}}\ell\sin\omega t \qquad z = \tfrac{1}{\sqrt{2}}\ell$$

According to eq.(7.18), the inertia tensor for m, there being only one mass point to sum over, is thus

$$\begin{aligned} I &= m\begin{pmatrix} r^2-x^2 & -xy & -xz \\ -xy & r^2-y^2 & -yz \\ -xz & -yz & r^2-z^2 \end{pmatrix} \\ &= m\begin{pmatrix} \ell^2 - \frac{1}{2}\ell^2\cos^2\omega t & -\frac{1}{2}\ell^2\cos\omega t\sin\omega t & -\frac{1}{2}\ell^2\cos\omega t \\ -\frac{1}{2}\ell^2\cos\omega t\sin\omega t & \ell^2 - \frac{1}{2}\ell^2\sin^2\omega t & -\frac{1}{2}\ell^2\sin\omega t \\ -\frac{1}{2}\ell^2\cos\omega t & -\frac{1}{2}\ell^2\sin\omega t & \ell^2 - \frac{1}{2}\ell^2 \end{pmatrix} \\ &= m\ell^2\begin{pmatrix} 1 - \frac{1}{2}\cos^2\omega t & -\frac{1}{2}\cos\omega t\sin\omega t & -\frac{1}{2}\cos\omega t \\ -\frac{1}{2}\cos\omega t\sin\omega t & 1 - \frac{1}{2}\sin^2\omega t & -\frac{1}{2}\sin\omega t \\ -\frac{1}{2}\cos\omega t & -\frac{1}{2}\sin\omega t & \frac{1}{2} \end{pmatrix} \end{aligned}$$

Since

$$\boldsymbol{\omega} = \omega\,\hat{\mathbf{z}} = \begin{pmatrix} 0 \\ 0 \\ \omega \end{pmatrix}$$

the angular momentum $\mathbf{L} = I\boldsymbol{\omega}$ works out to

$$\mathbf{L} = \tfrac{1}{2}m\omega\ell^2\begin{pmatrix} -\cos\omega t \\ -\sin\omega t \\ 1 \end{pmatrix}$$

which, as it has x and y as well as z components, is not parallel to $\boldsymbol{\omega}$. $\mathbf{L}$ also depends on time, which means that there is a nonzero torque involved:

$$\boldsymbol{\tau} = \frac{d\mathbf{L}}{dt} = \frac{d}{dt}\left[\tfrac{1}{2}m\omega\ell^2\begin{pmatrix} -\cos\omega t \\ -\sin\omega t \\ 1 \end{pmatrix}\right] = \tfrac{1}{2}m\omega^2\ell^2\begin{pmatrix} \sin\omega t \\ -\cos\omega t \\ 0 \end{pmatrix}$$

To make sense of this, consider the situation at $t = 0$: the mass is then at $(x, y, z) = (\frac{1}{\sqrt{2}}\ell, 0, \frac{1}{\sqrt{2}}\ell)$ — that is, is in the xz plane — and is moving in the

[11] We are not worrying about being general about initial conditions here, since the initial conditions are irrelevant to the effect in which we are interested.

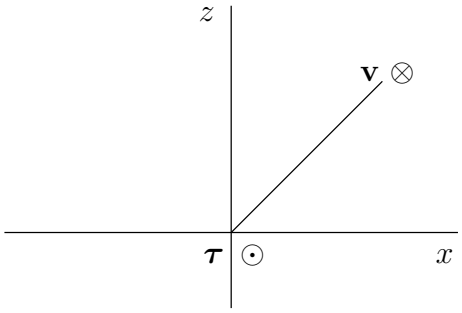


Figure 7.8: More Rotational Weirdness

positive y direction, as shown in fig. (7.8). At this moment

$$\boldsymbol{\tau} = \tfrac{1}{2} m\omega^2 \ell^2 \begin{pmatrix} 0 \\ -1 \\ 0 \end{pmatrix} = -\tfrac{1}{2} m\omega^2 \ell^2 \, \hat{\mathbf{y}}$$

By the right-hand rule, a torque coming out of the page in fig. (7.8) corresponds to a counterclockwise twist being applied to the mass m, which is just what we would expect: in the absence of such a torque, centrifugal effects would cause the mass to swing clockwise toward the x axis. If the mass is to maintain the rotation we specified for it, there must therefore be a counterclockwise torque to restrain it. Effects like this produce the weird behavior you sometimes see in mobiles and similar toys sometimes found on the desktops of business executives, doctors, lawyers, and other parasites on society.

7.3 Coriolis Effects

Suppose we have two reference frames that are aligned at time $t = 0$: a stationary frame and frame rotating at constant angular velocity ω about the z axis. At a later time t, when the rotating frame has rotated through an angle $\theta = \omega t$ relative to the stationary frame, the relationships between the axes are as shown in fig. (7.9), in which all of the marked angles are equal to θ. The coordinates (x, y, z) in the stationary frame are, as you can sort of see from fig. (7.9), related to the coordinates (x', y', z') in the rotating frame by

$$\begin{aligned} x' &= \cos\theta \, x + \sin\theta \, y \\ &= \cos\omega t \, x + \sin\omega t \, y \\ y' &= -\sin\theta \, x + \cos\theta \, y \\ &= -\sin\omega t \, x + \cos\omega t \, y \\ z' &= z \end{aligned}$$

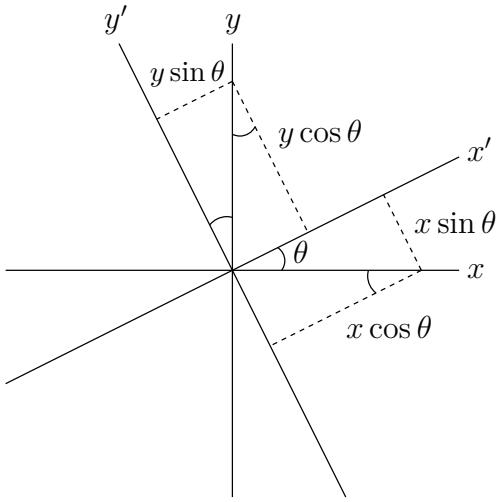


Figure 7.9: The Two Reference Frames

These relations can be written in matrix form as

$$\mathbf{r}' = R\mathbf{r} \tag{7.29}$$

where $\mathbf{r}$ and $\mathbf{r}'$ are the position vectors corresponding to (x, y, z) and (x', y', z'), and where R is the rotation matrix

$$R = \begin{pmatrix} \cos\omega t & \sin\omega t & 0 \\ -\sin\omega t & \cos\omega t & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The inverse R^{-1} of the matrix R is

$$R^{-1} = \begin{pmatrix} \cos\omega t & -\sin\omega t & 0 \\ \sin\omega t & \cos\omega t & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

as you can see either by doing out the matrix multiplication to verify that $RR^{-1} = R^{-1}R = 1$ or by noting that to undo a rotation by angle ωt, you simply do a rotation by angle $-\omega t$.[12] Multiplying both sides of eq. (7.29) by R^{-1}, we obtain

$$\mathbf{r} = R^{-1}\mathbf{r}' \tag{7.30}$$

Now, $\mathbf{F} = m\mathbf{a}$ holds only in inertial frames, which means that it will hold in the stationary frame but not, since rotation involves acceleration, in the rotating frame. We must therefore begin from Newton's second law in the stationary frame:

$$\mathbf{F}_{true} = m\mathbf{a} = m\ddot{\mathbf{r}} \tag{7.31}$$

[12]Note that R^{-1} is just the transpose R^t of R. $R^{-1} = R^t$ is a general property of matrix representations of the orthogonal (rotation) groups $O(n)$. Our matrix $R \in O(3)$.

where the subscript *true* is to indicate that this force is the true physical force (as opposed to the fictitious forces we will ultimately discover in the rotating frame). Our object is to determine what sort of force law is obeyed in the rotating frame.

To see what things look like in the rotating frame, we apply the rotation operator R to both sides of eq. (7.31):

$$R\mathbf{F}_{true} = mR\ddot{\mathbf{r}}$$

Now, $R\mathbf{F}_{true}$ is just the rotated form of the true physical force $\mathbf{F}_{true}$ and is therefore the true physical force $\mathbf{F}'_{true}$ in the rotating frame, so that this becomes

$$\mathbf{F}'_{true} = mR\ddot{\mathbf{r}}$$

The next step is to write the right-hand side $mR\ddot{\mathbf{r}}$ in terms of the position vector $\mathbf{r}'$ of the rotating frame rather than the position vector $\mathbf{r}$ of the stationary frame, which we accomplish by using eq. (7.30):

$$\mathbf{F}'_{true} = mR\frac{d^2}{dt^2}(R^{-1}\mathbf{r}')$$

Using the product rule to expand the d^2/dt^2, we have

$$\begin{aligned}\mathbf{F}'_{true} &= mR\left(\frac{d^2R^{-1}}{dt^2}\mathbf{r}' + 2\frac{dR^{-1}}{dt}\dot{\mathbf{r}}' + R^{-1}\ddot{\mathbf{r}}'\right)\\ &= mR\frac{d^2R^{-1}}{dt^2}\mathbf{r}' + 2mR\frac{dR^{-1}}{dt}\dot{\mathbf{r}}' + m\ddot{\mathbf{r}}' \qquad (7.32)\end{aligned}$$

The $m\ddot{\mathbf{r}}' = m\mathbf{a}'$ term is just what an observer in the rotating frame would report for mass times acceleration and therefore what the rotating observer might naively set equal to the net force by Newton's second law. Since Newton's second law doesn't actually hold in the rotating frame, we will label this force $\mathbf{F}'_{apparent}$:

$$\mathbf{F}'_{apparent} = m\mathbf{a}' = m\ddot{\mathbf{r}}'$$

Thus eq. (7.32) becomes

$$\mathbf{F}'_{true} = mR\frac{d^2R^{-1}}{dt^2}\mathbf{r}' + 2mR\frac{dR^{-1}}{dt}\dot{\mathbf{r}}' + \mathbf{F}'_{apparent}$$

or, if we solve this for $\mathbf{F}'_{apparent}$,

$$\mathbf{F}'_{apparent} = \mathbf{F}'_{true} - mR\frac{d^2R^{-1}}{dt^2}\mathbf{r}' - 2mR\frac{dR^{-1}}{dt}\dot{\mathbf{r}}' \qquad (7.33)$$

Our remaining task is to work out the last two terms on the right-hand side and then interpret our results.

If we take a time derivative of R^{-1}, we have

$$\begin{aligned}\frac{dR^{-1}}{dt} &= \frac{d}{dt}\begin{pmatrix}\cos\omega t & -\sin\omega t & 0\\ \sin\omega t & \cos\omega t & 0\\ 0 & 0 & 1\end{pmatrix}\\ &= \begin{pmatrix}-\omega\sin\omega t & -\omega\cos\omega t & 0\\ \omega\cos\omega t & -\omega\sin\omega t & 0\\ 0 & 0 & 0\end{pmatrix}\\ &= -\omega\begin{pmatrix}\sin\omega t & \cos\omega t & 0\\ -\cos\omega t & \sin\omega t & 0\\ 0 & 0 & 0\end{pmatrix}\end{aligned}$$

Taking a second time derivative, we have

$$\begin{aligned}\frac{d^2R^{-1}}{dt^2} &= -\omega\frac{d}{dt}\begin{pmatrix}\sin\omega t & \cos\omega t & 0\\ -\cos\omega t & \sin\omega t & 0\\ 0 & 0 & 0\end{pmatrix}\\ &= -\omega\begin{pmatrix}\omega\cos\omega t & -\omega\sin\omega t & 0\\ \omega\sin\omega t & \omega\cos\omega t & 0\\ 0 & 0 & 0\end{pmatrix}\\ &= -\omega^2\begin{pmatrix}\cos\omega t & -\sin\omega t & 0\\ \sin\omega t & \cos\omega t & 0\\ 0 & 0 & 0\end{pmatrix}\end{aligned}$$

Thus

$$\begin{aligned}R\frac{d^2R^{-1}}{dt^2} &= \begin{pmatrix}\cos\omega t & \sin\omega t & 0\\ -\sin\omega t & \cos\omega t & 0\\ 0 & 0 & 1\end{pmatrix}\left[-\omega^2\begin{pmatrix}\cos\omega t & -\sin\omega t & 0\\ \sin\omega t & \cos\omega t & 0\\ 0 & 0 & 0\end{pmatrix}\right]\\ &= -\omega^2\begin{pmatrix}\cos^2\omega t+\sin^2\omega t & -\cos\omega t\sin\omega t+\sin\omega t\cos\omega t & 0\\ -\sin\omega t\cos\omega t+\cos\omega t\sin\omega t & \sin^2\omega t+\cos^2\omega t & 0\\ 0 & 0 & 0\end{pmatrix}\\ &= -\omega^2\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\end{pmatrix}\end{aligned}$$

and likewise

$$R\frac{dR^{-1}}{dt} = \begin{pmatrix}\cos\omega t & \sin\omega t & 0\\ -\sin\omega t & \cos\omega t & 0\\ 0 & 0 & 1\end{pmatrix}\left[-\omega\begin{pmatrix}\sin\omega t & \cos\omega t & 0\\ -\cos\omega t & \sin\omega t & 0\\ 0 & 0 & 0\end{pmatrix}\right]$$

$$= -\omega \begin{pmatrix} \cos\omega t \sin\omega t - \sin\omega t \cos\omega t & \cos^2\omega t + \sin^2\omega t & 0 \\ -\sin^2\omega t - \cos^2\omega t & -\sin\omega t\cos\omega t + \cos\omega t \sin\omega t & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

$$= -\omega \begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Using these results in eq. (7.33), we have

$$\begin{aligned} \mathbf{F}'_{apparent} &= \mathbf{F}'_{true} - mR\frac{d^2R^{-1}}{dt^2}\mathbf{r}' - 2mR\frac{dR^{-1}}{dt}\dot{\mathbf{r}}' \\ &= \mathbf{F}'_{true} - m\left[-\omega^2\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}\right]\mathbf{r}' \\ &\qquad - 2m\left[-\omega\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\right]\dot{\mathbf{r}}' \\ &= \mathbf{F}'_{true} + m\omega^2\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} \\ &\qquad + 2m\omega\begin{pmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \dot{x}' \\ \dot{y}' \\ \dot{z}' \end{pmatrix} \\ &= \mathbf{F}'_{true} + m\omega^2\begin{pmatrix} x' \\ y' \\ 0 \end{pmatrix} + 2m\omega\begin{pmatrix} \dot{y}' \\ -\dot{x}' \\ 0 \end{pmatrix} \end{aligned}$$

Now,

$$\begin{pmatrix} x' \\ y' \\ 0 \end{pmatrix}$$

is just $\mathbf{r}'$ with its z component lopped off — that is, it is the part of $\mathbf{r}'$ that is perpendicular to the z axis, which is the axis of rotation. We can therefore write the middle term above as

$$m\omega^2\mathbf{r}'_\perp$$

To see what the third term corresponds to, consider

$$\boldsymbol{\omega} \times \mathbf{v}' = \omega\,\widehat{\mathbf{z}}' \times \mathbf{v}' = \begin{vmatrix} \widehat{\mathbf{x}}' & \widehat{\mathbf{y}}' & \widehat{\mathbf{z}}' \\ 0 & 0 & \omega \\ v'_x & v'_y & v'_z \end{vmatrix} = -\omega v'_y\,\widehat{\mathbf{x}}' + \omega v'_x\,\widehat{\mathbf{y}}'$$

which, written as a column vector, is

$$\boldsymbol{\omega} \times \mathbf{v}' = \omega \begin{pmatrix} -v'_y \\ v'_x \\ 0 \end{pmatrix} = \omega \begin{pmatrix} -\dot{y}' \\ \dot{x}' \\ 0 \end{pmatrix}$$

The third term in our result for $\mathbf{F}'_{apparent}$ may thus be expressed as

$$2m\omega \begin{pmatrix} \dot{y}' \\ -\dot{x}' \\ 0 \end{pmatrix} = -2m\boldsymbol{\omega} \times \mathbf{v}'$$

Putting this all together, we finally arrive at

$$\mathbf{F}'_{apparent} = \mathbf{F}'_{true} + m\omega^2 \mathbf{r}'_\perp - 2m\boldsymbol{\omega} \times \mathbf{v}' \tag{7.34}$$

In addition to the rotated form $\mathbf{F}'_{true}$ of the true physical force $\mathbf{F}_{true}$, in the rotating frame there are also two fictitious forces — forces that seem to observers in the rotating frame to be acting, but which are in fact only artifacts of the their frame's rotation and not true physical forces at all. The $m\omega^2\mathbf{r}'_\perp$ is just the familiar centrifugal force: it is in the outward radial direction and of magnitude $m\omega^2 r'_\perp$. The $-2m\boldsymbol{\omega} \times \mathbf{v}'$ is probably new to you: it is the Coriolis force.

The Coriolis force corresponds to your tangential velocity $\omega r'_\perp$ getting "out of sync" when you move you change your distance $r'_\perp$ from the axis of rotation. To see this, consider the simple case that the rotating frame is an LP record [13] and imagine that a roach that was sitting on and thus rotating with the record gets bored, takes off, and tries to fly [14] (from its perspective in the rotating frame) radially outward, as shown by the black line in fig. (7.10). As the roach moves out to larger $r_\perp$, the tangential velocity $\omega r_\perp$ that it shared with the record when it took off falls farther and farther behind the tangential velocity of the part of the record underneath it. As a result, the roach "loses ground" in the tangential direction and falls rotationally behind the record, as shown by the blue line in fig. (7.10). In eq. (7.34), the $\boldsymbol{\omega}$ and the $\mathbf{v}'$ in $-2m\boldsymbol{\omega} \times \mathbf{v}'$ are in this case $\odot$ and $\rightarrow$, respectively, so that $-2m\boldsymbol{\omega} \times \mathbf{v}'$ is indeed $\downarrow$.

[13] If you don't know what an LP record is, ask your parents. They will be amused. Or possibly depressed.

[14] Yes, those big roaches you see in all the buildings on campus can fly, though only for short distances and only when they are very agitated. So be nice to them.

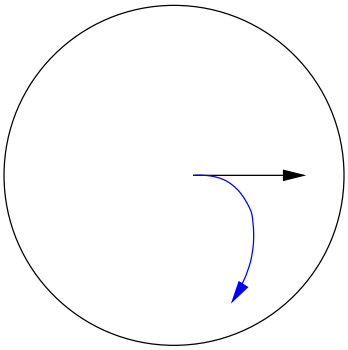


Figure 7.10: The Coriolis Effect

On our rotating Earth, the Coriolis effect is responsible for the rotations of weather systems. The cold, dry air of high-pressure systems is relatively dense, so that it descends and tries to spread outward along the surface of the Earth as shown by the black arrows on the left side of fig. (7.11). In the northern hemisphere, the air that is moving to the north, as it follows the Earth's curved surface, is moving to smaller $r_\perp$, so that its tangential velocity is "too big" and gets ahead of the Earth's rotation, as shown by the blue arrow. In eq. (7.34), if we make the Earth's axis the z axis, $\boldsymbol{\omega}$ is in the $\hat{\mathbf{z}}$ direction and the part of $\mathbf{v}'$ perpendicular to $\boldsymbol{\omega}$ is going into the page ($\otimes$), so that $-2m\boldsymbol{\omega} \times \mathbf{v}'$ is to the east. Air moving to the south experiences just the opposite effect. The overall result is that high-pressure systems rotate clockwise in the northern hemisphere. For low-pressure systems, it is just the opposite: their warm, moist air is relatively light, so that it rises and draws itself inward along the surface of the Earth as shown by the black arrows on the right side of fig. (7.11). The Coriolis effect thus goes the opposite way for low-pressure systems, which rotate counterclockwise in the

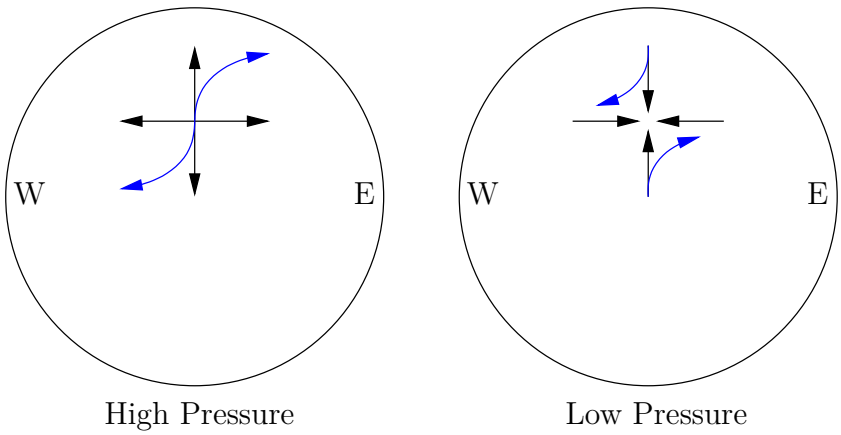


Figure 7.11: The Rotation of Weather Systems

northern hemisphere. And of course the rotational directions of both high- and low-pressure systems are reversed in the southern hemisphere, since in the southern hemisphere $r_\perp$ decreases as one moves southward.

In spite of what you may have seen on *The Simpsons*, the Coriolis effect is not responsible for the direction in which water swirls down drains: while significant for weather systems, the Coriolis effect is negligible on the scale of household plumbing, as you will actually show in problem #45. What causes the water to swirl down drains is a combination of asymmetries in the plumbing and conservation of angular momentum — even though the water in the sink may look still, an imperceptible swirl at large radii will become quite significant at small radii.

And now back to reality ...

7.4 Constant Angular Acceleration

By analogy, our constant-acceleration relations of §3.2.1 for translational motion,

$$\begin{aligned}
v &= v_0 + at & & & \omega &= \omega_0 + \alpha t && (7.35a)\\
x - x_0 &= v_0 t + \tfrac{1}{2}at^2 & & & \theta - \theta_0 &= \omega_0 t + \tfrac{1}{2}\alpha t^2 && (7.35b)\\
&= vt - \tfrac{1}{2}at^2 & \text{become} & & &= \omega t - \tfrac{1}{2}\alpha t^2 && (7.35c)\\
&= \frac{v_0 + v}{2}\, t & & & &= \frac{\omega_0 + \omega}{2}\, t && (7.35d)\\
v^2 - v_0^2 &= 2a(x - x_0) & & & \omega^2 - \omega_0^2 &= 2\alpha(\theta - \theta_0) && (7.35e)
\end{aligned}$$

Since you already know how to use these relations, that's about all we'll be saying on the subject.

Okay, maybe just one example. Consider that innocent playground diversion known as a merry-go-round,[15] onto the perimeter of which the smallest, weakest kid in the playground is secured and then, through the combined work of hordes of other kids, all foaming at the mouth with blood lust, accelerated from rest up to some final angular velocity ω_{death} — large enough

[15]Not to be confused with a carousel, which sometimes is also known as a merry-go-round: we mean the spinning-disk-with-bars-to-hold-onto-thingie frequently found between the seesaws and the monkey bars; the carousel is the fair-ground ride with the little horsies on it, the one where you try to grab the brass ring. Although some of your generation, thanks to the efforts of lawyers seeking to ensure that there is no human activity without legal liability, may be totally unfamiliar with merry-go-rounds, not to mention seesaws and monkey bars.

to cause loss of consciousness — in time t_{death}.[16] Under the assumption that the angular acceleration of the merry-go-round is constant, to determine the number of revolutions made by the merry-go-round during its acceleration our first step would be to figure out the angle through which it rotated using eq. (7.35d):

$$\theta - \theta_0 = \frac{\omega_0 + \omega}{2}\, t = \tfrac{1}{2}\omega_{death} t_{death}$$

Since there is one revolution for every 2π radians, the corresponding number of revolutions is thus

$$\text{number of revolutions} = \frac{\theta - \theta_0}{2\pi} = \frac{\omega_{death} t_{death}}{4\pi}$$

7.5 Moments of Inertia

Table (7.2) lists the moments of inertia about perpendicular axes through some of the more common symmetric distributions of mass.

Shape	Moment of Inertia
Point mass	mr^2
Thin ring or cylindrical shell	mr^2
Solid disk or cylinder	$\frac{1}{2}mr^2$
Spherical shell	$\frac{2}{3}mr^2$
Solid sphere	$\frac{2}{5}mr^2$
Thin rod (about end)	$\frac{1}{3}m\ell^2$
Thin rod (about center)	$\frac{1}{12}m\ell^2$

Table 7.2: Moments of Inertia

We will now derive these results in painful detail by applying eqq. (7.3) and (7.4).

Point mass: In this case, there is only a single contribution to the sum

$$I = \sum m_i r_{\perp,i}^2$$

so we get simply $I = mr^2$, where r is the radial distance of the point mass from the axis of rotation.

[16] This was, of course, the principal entertainment provided by merry-go-rounds, just as that of the seesaw was to see whether you could seesaw violently enough to throw the guy on the other end off, that of the swings was to see how close you could get to horizontal while standing in the seat, and that of those little animals on springs was to see how close you could come to hitting the ground. Ah, those were the days. Your generation doesn't know what it's missing. Sad, really.

Cylindrical shell or thin ring rotated about its axis: For the thin ring of radius r, all the mass is at the same perpendicular distance r from the axis of rotation, so that $r_\perp = r$, being constant, may be pulled outside of the integration:

$$I = \int dm\, r_\perp^2 = \left(\int dm\right) r^2 = mr^2$$

Rotationally, the thin ring is therefore no different from the point mass: in both cases, all of the mass is at the same perpendicular distance from the axis of rotation. And since stretching the mass of the ring into a cylindrical shell still leaves all the mass at the same perpendicular distance r from the axis of rotation, the moment of inertia of a cylindrical shell will also be mr^2.

Solid cylinder or disk rotated about its axis: If the disk is uniform, the mass m of the disk is evenly spread over its surface area πR^2, so that it has a surface mass density (mass per unit area) of $\sigma = m/\pi R^2$. (Note that we are using R for the radius of the disk; we will need r to represent a variable radial distance that can span from $r = 0$ at the center of the disk to $r = R$ at its perimeter.) The mass dm of a patch of area dA is therefore

$$dm = \sigma\, dA = \frac{m}{\pi R^2}\, dA$$

and we need to integrate over the area of the disk to get I:

$$I = \int dm\, r_\perp^2 = \int\limits_{\text{disk}} \frac{m}{\pi R^2}\, dA\, r^2$$

To integrate over a circular disk, polar coordinates are most natural. In polar coordinates, the area element dA is $r\, dr\, d\theta$, so our integration over the disk becomes

$$\begin{aligned} I &= \int_0^R r\, dr \int_0^{2\pi} d\theta\, \frac{m}{\pi R^2}\, r^2 \\ &= \frac{m}{\pi R^2} \int_0^R r^3\, dr \int_0^{2\pi} d\theta \\ &= \frac{m}{\pi R^2} \Big[\tfrac{1}{4} r^4\Big]_0^R \Big[\theta\Big]_0^{2\pi} \\ &= \frac{m}{\pi R^2} (\tfrac{1}{4} R^4)(2\pi) \\ &= \tfrac{1}{2} m R^2 \end{aligned}$$

Again, since stretching the mass of the disk into a cylinder does not change the perpendicular distance of the mass from the axis of rotation, the moment of inertia of a cylinder will also be $\frac{1}{2} mR^2$.

Spherical shell rotated about an axis through its center: If the shell is uniform, the mass m of the shell is evenly spread over its surface area $4\pi R^2$, so that it has a surface mass density of $\sigma = m/4\pi R^2$. The mass dm of a patch of area dA is therefore

$$dm = \sigma\, dA = \frac{m}{4\pi R^2}\, dA$$

and we need to integrate over the area of the shell to get I:

$$I = \int dm\; r_\perp^2 = \int\limits_{\text{shell}} \frac{m}{4\pi R^2}\, dA\; r_\perp^2$$

To integrate over a sphere, spherical coordinates are most natural. On the surface of the shell, where $r = R$, the area element dA is $r^2 \sin\theta\, d\theta\, d\varphi = R^2 \sin\theta\, d\theta\, d\varphi$. And if we take the z axis to be the axis of rotation, $r_\perp$ is the radius in the plane perpendicular to the z axis, that is, $r_\perp = r\sin\theta = R\sin\theta$.[17] Our integration over the shell thus becomes

$$\begin{aligned} I &= \int_0^\pi d\theta \int_0^{2\pi} d\varphi\; R^2 \sin\theta\; \frac{m}{4\pi R^2}\; (R\sin\theta)^2 \\ &= \frac{m}{4\pi R^2}\; R^4 \int_0^\pi d\theta\; \sin^3\theta \int_0^{2\pi} d\varphi \\ &= \frac{m}{4\pi R^2}\; R^4 \int_{-1}^1 d(\cos\theta)\; (1-\cos^2\theta)\; (2\pi) \\ &= \frac{m}{4\pi R^2}\; R^4 \Big[\cos\theta - \tfrac{1}{3}\cos^3\theta\Big]_{-1}^1\; (2\pi) \\ &= \frac{m}{4\pi R^2}\; R^4\; (\tfrac{4}{3})\; (2\pi) \\ &= \tfrac{2}{3} m R^2 \end{aligned}$$

Solid sphere rotated about an axis through its center: If the sphere is uniform, the mass m of the sphere is evenly spread throughout its volume $\frac{4}{3}\pi R^3$, so that it has a volume mass density of $\rho = m/\frac{4}{3}\pi R^3$. The mass dm of a volume element dV is therefore

$$dm = \rho\, dV = \frac{m}{\frac{4}{3}\pi R^3}\, dV$$

[17] If this isn't immediately clear, regard $r_\perp$ as the perpendicular distance from the z axis given by

$$r_\perp = \sqrt{x^2 + y^2} = \sqrt{(r\sin\theta\cos\varphi)^2 + (r\sin\theta\sin\varphi)^2} = \sqrt{r^2\sin^2\theta(\cos^2\varphi + \sin^2\varphi)} = r\sin\theta$$

and we need to integrate over the volume of the sphere to get I:

$$I = \int dm\, r_\perp^2 = \int\limits_{\text{sphere}} \frac{m}{\frac{4}{3}\pi R^3}\, dV\, r_\perp^2$$

Spherical coordinates are again most natural. This time, however, we will be integrating over r as well as over θ and φ: our volume element is $dV = r^2 \sin\theta\, dr\, d\theta\, d\varphi$. And if we again take the z axis to be the axis of rotation, $r_\perp = r\sin\theta$. Our integration over the spherical volume thus becomes

$$\begin{aligned} I &= \int_0^R dr \int_0^\pi d\theta \int_0^{2\pi} d\varphi\, r^2 \sin\theta\, \frac{m}{\frac{4}{3}\pi R^3}\, (r\sin\theta)^2 \\ &= \frac{m}{\frac{4}{3}\pi R^3} \int_0^R dr\, r^4 \int_0^\pi d\theta\, \sin^3\theta \int_0^{2\pi} d\varphi \\ &= \frac{m}{\frac{4}{3}\pi R^3}\, (\tfrac{1}{5}R^5)\, (\tfrac{4}{3})\, (2\pi) \\ &= \tfrac{2}{5}mR^2 \end{aligned}$$

Thin rod about its end: For the case of a thin rod rotated about a perpendicular axis through one of its ends, we need to carry out the integration

$$I = \int dm\, r_\perp^2$$

over the rod. If the mass m of the rod is uniformly distributed along its length ℓ, it has a linear mass density (mass per unit length) of $\lambda = m/\ell$, so that the mass dm of a segment of length dx is

$$dm = \lambda\, dx = \frac{m}{\ell}\, dx$$

Integrating over the length of the rod, we have

$$I = \int dm\, r^2 = \int_0^\ell \frac{m}{\ell}\, dx\, x^2 = \frac{m}{\ell}\tfrac{1}{3}\ell^3 = \tfrac{1}{3}m\ell^2$$

Thin rod about its center: There are two ways to arrive at the result for I when the perpendicular axis is through the center rather than through an end of the rod. First, we can carry out the above integration, but with x going from $-\frac{1}{2}\ell$ to $+\frac{1}{2}\ell$ rather than from 0 to ℓ:

$$I = \frac{m}{\ell}\tfrac{1}{3}x^3\Big|_{-\frac{1}{2}\ell}^{\frac{1}{2}\ell} = \tfrac{1}{12}m\ell^2$$

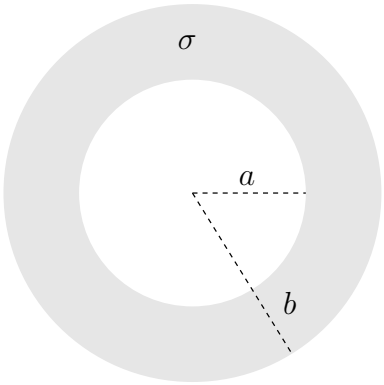


Figure 7.12: Look out! An Annulus!

Alternatively, we can regard the rod rotated about its center as two half-rods rotated about their ends. Each of these half-rods will have a mass $\frac{1}{2}m$, a length $\frac{1}{2}\ell$, and consequently a moment of inertia of

$$\frac{1}{3}\frac{m}{2}\left(\frac{\ell}{2}\right)^2 = \tfrac{1}{24}m\ell^2$$

The moment of inertia of the whole rod will be just twice this, which gives the same result as the integration.

This alternative method is a special case of a general property that moments of inertia share with centers of mass, the property of superposition: you can get a result for the moment of inertia of a "compound" object simply by adding up the moments of inertia of its various parts. Similarly, you can deal with holes in objects by subtracting out the moment of inertia that the hole would have had.

For example, suppose you have an annulus[18] of uniform surface density σ and inner and outer radii a and b, as shown in fig. (7.12). You can get the moment of inertia of the annulus just by subtracting the moment of inertia of a solid disk of radius a from that of a solid disk of radius b. The mass of a solid disk of radius b that has uniform density σ would be $m_b = \sigma\pi b^2$; that of a solid disk of radius a would be $m_a = \sigma\pi a^2$. The moment of inertia of the annulus is therefore

$$I = \tfrac{1}{2}m_b b^2 - \tfrac{1}{2}m_a a^2 = \tfrac{1}{2}(\sigma\pi b^2)b^2 - \tfrac{1}{2}(\sigma\pi a^2)a^2 = \tfrac{1}{2}\sigma\pi(b^4 - a^4)$$

[18] An annulus is of course just a washer. But it's a much cooler word.

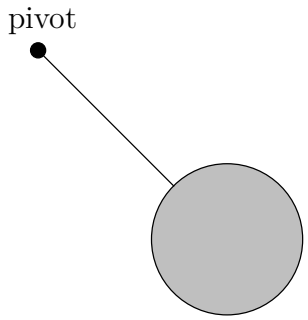


Figure 7.13: Nasty — Not!

7.6 The Parallel-Axis Theorem

The parallel-axis theorem enables us to determine the moment of inertia of a body about any axis in terms of its moment of inertia about a parallel axis through its center of mass. Using it we could, for example, figure out the moment of inertia of the pendulum disk of fig. (7.13) about the pivot point without doing a nasty $\int dm\, r_\perp^2$ integration over the disk.

Recall once again (from eq. (6.10) on p.303 and the accompanying figure) that we may re-express the position vector $\mathbf{r}_i$ of mass element m_i in terms of the position vector $\mathbf{r}_{cm}$ of the center of mass and the position vector $\mathbf{r}_{i,cm}$ of m_i relative to the center of mass:

$$\mathbf{r}_i = \mathbf{r}_{cm} + \mathbf{r}_{i,cm} \tag{7.36}$$

For a planar or effectively planar distribution of mass in the xy plane, the moment of inertia about the z axis will be

$$I = \sum_{i=1}^{n} m_i\, r_{i\perp}^2 = \sum_{i=1}^{n} m_i\, r_i^2$$

which, if we use eq. (7.36), becomes

$$\begin{aligned}
I &= \sum_{i=1}^{n} m_i\, r_i^2 \\
&= \sum_{i=1}^{n} m_i\, (\mathbf{r}_{cm} + \mathbf{r}_{i,cm})^2 \\
&= \sum_{i=1}^{n} m_i\, r_{cm}^2 + \sum_{i=1}^{n} m_i\, 2\mathbf{r}_{cm} \cdot \mathbf{r}_{i,cm} + \sum_{i=1}^{n} m_i\, r_{i,cm}^2 \\
&= \left(\sum_{i=1}^{n} m_i\right) r_{cm}^2 + 2\mathbf{r}_{cm} \cdot \left(\sum_{i=1}^{n} m_i\, \mathbf{r}_{i,cm}\right) + \sum_{i=1}^{n} m_i r_{i,cm}^2 \\
&= M r_{cm}^2 + 2\mathbf{r}_{cm} \cdot M\left(\frac{1}{M}\sum_{i=1}^{n} m_i\, \mathbf{r}_{i,cm}\right) + \sum_{i=1}^{n} m_i r_{i,cm}^2
\end{aligned}$$

The parenthetic expression in the middle term is the position vector of the center of mass relative to the center of mass and therefore vanishes. We are then left with

$$I = Mr_{cm}^2 + \sum_{i=1}^{n} m_i r_{i,cm}^2$$

The first term is the result we would obtain for I if we collapsed the body down to a single point at the center of mass. The second term is the expression by which we would determine I about the axis, were it through the center of mass. Thus we arrive at the parallel-axis theorem:

$$I = I_{about\ cm} + mR^2 \tag{7.37}$$

where $I_{about\ cm}$ is the moment of inertia about a parallel axis through the center of mass, and R is the perpendicular distance from the axis about which we are evaluating I and the parallel axis passing through the center of mass. That might sound involved, but it's not: to determine the moment of inertia of a body about an axis off its center of mass, all you have to do is take the moment of inertia you would have had if the axis had passed through the center of mass, and then add to it a point-mass-like contribution mR^2, where m is the body's total mass and R is the perpendicular distance between the actual axis of rotation and the body's center of mass.

For a uniform thin rod, the moment of inertia about an axis that passes perpendicularly through its center is $\frac{1}{12}m\ell^2$. Since the center of mass of the rod coincides with its geometric center, the distance from the rod's center of mass to one of its ends is $\frac{1}{2}\ell$, and its moment of inertia about a parallel axis that passes perpendicularly through one of its ends is therefore

$$I_{about\ end} = I_{cm} + mR^2 = \tfrac{1}{12}m\ell^2 + m\left(\tfrac{1}{2}\ell\right)^2 = \tfrac{1}{3}m\ell^2 \quad \checkmark$$

There were, of course, other equally simple ways to obtain this result for the thin rod, but the parallel-axis theorem makes it easy to determine many otherwise nasty moments of inertia. If, for example, the disk shown in fig. (7.13) is uniform, of mass m and radius a, and the rod connecting it to the pivot massless and of length ℓ, then to calculate the moment of inertia about the pivot point we need simply note that $I_{about\ cm}$ will be the moment of inertia of the disk about its center, which is just $\frac{1}{2}ma^2$. The center of mass of the disk being a distance $\ell + a$ from the pivot, the moment of inertia about the pivot is thus

$$I = \tfrac{1}{2}ma^2 + m(\ell + a)^2$$

7.7 Conservation of Angular Momentum

Recall (eq. (6.9) on p.301) that the rate of change of a system's total linear momentum is determined by the net external force on the system:

$$\mathbf{F}_{net\ ext,\ system} = \frac{d\mathbf{p}_{total,\ system}}{dt}$$

We will now establish that a similar relation holds for angular momentum and torque. Recall (eq. (7.12) or (7.22)) that

$$\boldsymbol{\tau} = \frac{d\mathbf{L}}{dt} \tag{7.38}$$

If we sum over a system of mass points, this becomes

$$\begin{aligned}\sum_{i=1}^{n} \boldsymbol{\tau}_i &= \sum_{i=1}^{n} \frac{d\mathbf{L}_i}{dt} \\ &= \frac{d}{dt}\left(\sum_{i=1}^{n} \mathbf{L}_i\right) \\ \boldsymbol{\tau}_{net,\ system} &= \frac{d\mathbf{L}_{total,\ system}}{dt}\end{aligned} \tag{7.39}$$

Now, the individual torques

$$\boldsymbol{\tau}_i = \mathbf{r}_i \times \mathbf{F}_i$$

that make up the net torque on the system can be separated into torques due to external forces exerted on the system by sources outside the system and torques due to internal forces exerted by the various parts of the system on each other. As was the case with contributions to the net force, the contributions to the net torque that come from internal forces will cancel each other out pairwise by Newton's third law: for each pair ij of mass points in the system, if $\mathbf{F}_{ij}$ is the force exerted on m_i by m_j and $\mathbf{F}_{ji}$ that exerted on m_j by m_i, then the third law requires $\mathbf{F}_{ji} = -\mathbf{F}_{ij}$, and the pair of contributions to the net torque will be

$$\begin{aligned}\mathbf{r}_i \times \mathbf{F}_{ij} + \mathbf{r}_j \times \mathbf{F}_{ji} &= \mathbf{r}_i \times \mathbf{F}_{ij} - \mathbf{r}_j \times \mathbf{F}_{ij} \\ &= (\mathbf{r}_i - \mathbf{r}_j) \times \mathbf{F}_{ij}\end{aligned}$$

Since the vectors $\mathbf{r}_i - \mathbf{r}_j$ and $\mathbf{F}_{ij}$ are both along the line connecting m_i and m_j, their cross product vanishes, and we do not get any contribution to the net torque.[19]

[19] Actually, this argument is not valid for magnetic forces, which are not along the line connecting m_i and m_j, but the proof can be extended to include magnetic forces as well.

Thus the net torque on the system is due entirely to the torques of external forces, and eq. (7.39) becomes

$$\boldsymbol{\tau}_{\textit{net ext, system}} = \frac{d\mathbf{L}_{\textit{total, system}}}{dt} \tag{7.40}$$

So just as linear momentum is conserved when there is no net external force on a system, angular momentum is conserved when there is no net external torque on a system.

We will work through a couple of detailed examples of conservation of angular momentum later in this section, but first we want to note a useful decomposition of angular momentum into center-of-mass and relative pieces, similar to the decomposition (6.13a) of p.305 for kinetic energy.

For collections of mass points or extended bodies, the total angular momentum and torque will be

$$\mathbf{L}_{\textit{total, system}} = \sum_{i=1}^{n} \mathbf{r}_i \times \mathbf{p}_i \qquad \text{or} \qquad \int_{\text{body}} \mathbf{r} \times dm\, \mathbf{v} \tag{7.41}$$

$$\boldsymbol{\tau}_{\textit{total, system}} = \sum_{i=1}^{n} \mathbf{r}_i \times \mathbf{F}_i \qquad \text{or} \qquad \int_{\text{body}} \mathbf{r} \times dm\, \mathbf{a} \tag{7.42}$$

Recall that (eq. (6.10) on p.303 and the accompanying figure) we may re-express the position vector $\mathbf{r}_i$ of mass point m_i in terms of the position vector $\mathbf{r}_{cm}$ of the center of mass and the position vector $\mathbf{r}_{i,cm}$ of m_i relative to the center of mass:

$$\mathbf{r}_i = \mathbf{r}_{cm} + \mathbf{r}_{i,cm}$$

and thus, taking a time derivative,

$$\mathbf{v}_i = \mathbf{v}_{cm} + \mathbf{v}_{i,cm}$$

From eq. (7.41), we then have

$$\begin{aligned}
\mathbf{L}_{\textit{total, system}} &= \sum_{i=1}^{n} \mathbf{r}_i \times \mathbf{p}_i \\
&= \sum_{i=1}^{n} \mathbf{r}_i \times m_i \mathbf{v}_i \\
&= \sum_{i=1}^{n} (\mathbf{r}_{cm} + \mathbf{r}_{i,cm}) \times m_i (\mathbf{v}_{cm} + \mathbf{v}_{i,cm}) \\
&= \sum_{i=1}^{n} \mathbf{r}_{cm} \times m_i \mathbf{v}_{cm} + \sum_{i=1}^{n} \mathbf{r}_{i,cm} \times m_i \mathbf{v}_{cm} \\
&\qquad + \sum_{i=1}^{n} \mathbf{r}_{cm} \times m_i \mathbf{v}_{i,cm} + \sum_{i=1}^{n} \mathbf{r}_{i,cm} \times m_i \mathbf{v}_{i,cm}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{r}_{cm} \times \left(\sum_{i=1}^{n} m_i \right) \mathbf{v}_{cm} + \left(\sum_{i=1}^{n} m_i \mathbf{r}_{i,cm} \right) \times \mathbf{v}_{cm} \\
&\qquad + \ \mathbf{r}_{cm} \times \left(\sum_{i=1}^{n} m_i \mathbf{v}_{i,cm} \right) + \sum_{i=1}^{n} \mathbf{r}_{i,cm} \times m_i \mathbf{v}_{i,cm} \\
&= \mathbf{r}_{cm} \times M \mathbf{v}_{cm} + M \left(\frac{1}{M} \sum_{i=1}^{n} m_i \mathbf{r}_{i,cm} \right) \times \mathbf{v}_{cm} \\
&\qquad + \ \mathbf{r}_{cm} \times M \left(\frac{1}{M} \sum_{i=1}^{n} m_i \mathbf{v}_{i,cm} \right) + \sum_{i=1}^{n} \mathbf{r}_{i,cm} \times m_i \mathbf{v}_{i,cm}
\end{aligned}
$$

The parenthetic sum in the second term will give us the position vector of the center of mass *relative to the center of mass*, which is of course zero. The parenthetic sum in the third will will likewise give us the velocity of the center of mass *relative to the center of mass*, which is again zero. Thus we have

$$\mathbf{L}_{\textit{total, system}} = \mathbf{r}_{cm} \times M \mathbf{v}_{cm} + \sum_{i=1}^{n} \mathbf{r}_{i,cm} \times m_i \mathbf{v}_{i,cm}$$

The first term is what we would have for the angular momentum if the system were collapsed down to a single point at the center of mass, and we will therefore refer to it as the angular momentum of the center of mass and denote it by $\mathbf{L}_{\textit{of cm}}$:

$$\mathbf{L}_{\textit{of cm}} = \mathbf{r}_{cm} \times M \mathbf{v}_{cm} = \mathbf{r}_{cm} \times \mathbf{p}_{\textit{total, system}} \tag{7.43}$$

The second term is the sum of the torques on the system, *all evaluated relative to the center of mass*. We will therefore (though this is not a technical term) refer to it as the angular momentum about the center of mass and denote it by $\mathbf{L}_{\textit{about cm}}$:

$$\mathbf{L}_{\textit{about cm}} = \sum_{i=1}^{n} \mathbf{r}_{i,cm} \times m_i \mathbf{v}_{i,cm} \tag{7.44}$$

The total angular momentum of a system thus separates into a contribution from the center-of-mass motion and a contribution from the motion about the center of mass:

$$\mathbf{L}_{\textit{total, system}} = \mathbf{L}_{\textit{of cm}} + \mathbf{L}_{\textit{about cm}} \tag{7.45}$$

The total angular momentum of the Earth, for example, would be the sum of the angular momentum due to its orbit around the Sun ($\mathbf{L}_{\textit{of cm}}$) and the angular momentum due to its rotation about its own axis ($\mathbf{L}_{\textit{about cm}}$).

One hugely important related point is that

> *If there is no net external force on a system, it can rotate only about its center of mass.*

To see this, recall from eq. (6.7) that

$$\mathbf{F}_{net\ ext,\ system} = M\mathbf{a}_{cm} = M\frac{d\mathbf{v}_{cm}}{dt} \tag{7.46}$$

so that $\mathbf{v}_{cm}$ is constant when there is no net external force on the system. If the center of mass were experiencing any sort of rotation, its velocity would not be constant, so that any rotation that does occur must be about the center of mass. Stated another way, unless a body is constrained to rotate about some other point, it will naturally rotate about its center of mass; if it were to rotate about any other point, the rotation would cause the center of mass to revolve in a circular path, and this would of course require a centripetal force to be applied to the body (usually through the pivot point) by some external agent.

Eq. (7.46) also tells us what happens when there is a net external force on the system: that net external force determines the motion of the center of mass. So when you wing a ballpeen hammer at someone, in the air the center of mass of the hammer will (modulo air resistance) follow a nice parabolic arc due to the external gravitational force acting on it, whether the hammer is rotating or not. If the hammer is rotating, this rotation around the center of mass will simply be superposed on the parabolic trajectory, as shown in fig. (7.14). Depending on how rapidly the hammer is rotating, the head of

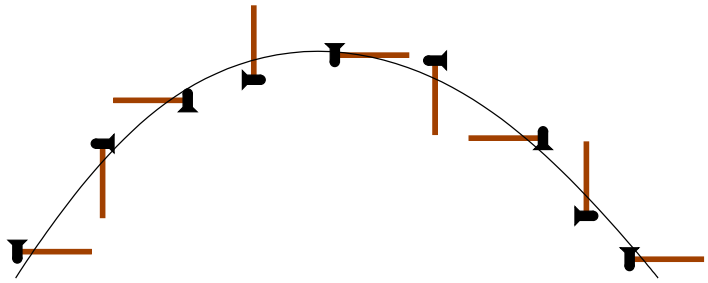


Figure 7.14: Excitement Is in the Air!

the hammer may trace out a curlicued path like that shown at the top of fig. (7.15), the more funky path shown in the middle of the figure, or the rather lame path shown at its bottom.

With a suitable net external force the center of mass of a system can of course be made to undergo any sort of motion, including rotational. If you nail a meter stick up through, say, the 20 cm mark and let it swing about that point, its center of mass (the center of the stick) will then swing in a circular arc, with the required centripetal force, supplied by the nail, being exerted on the meter stick at the pivot point: the nail will pull the center of mass of the stick radially inward toward the pivot point with the force needed for the center of mass's circular motion

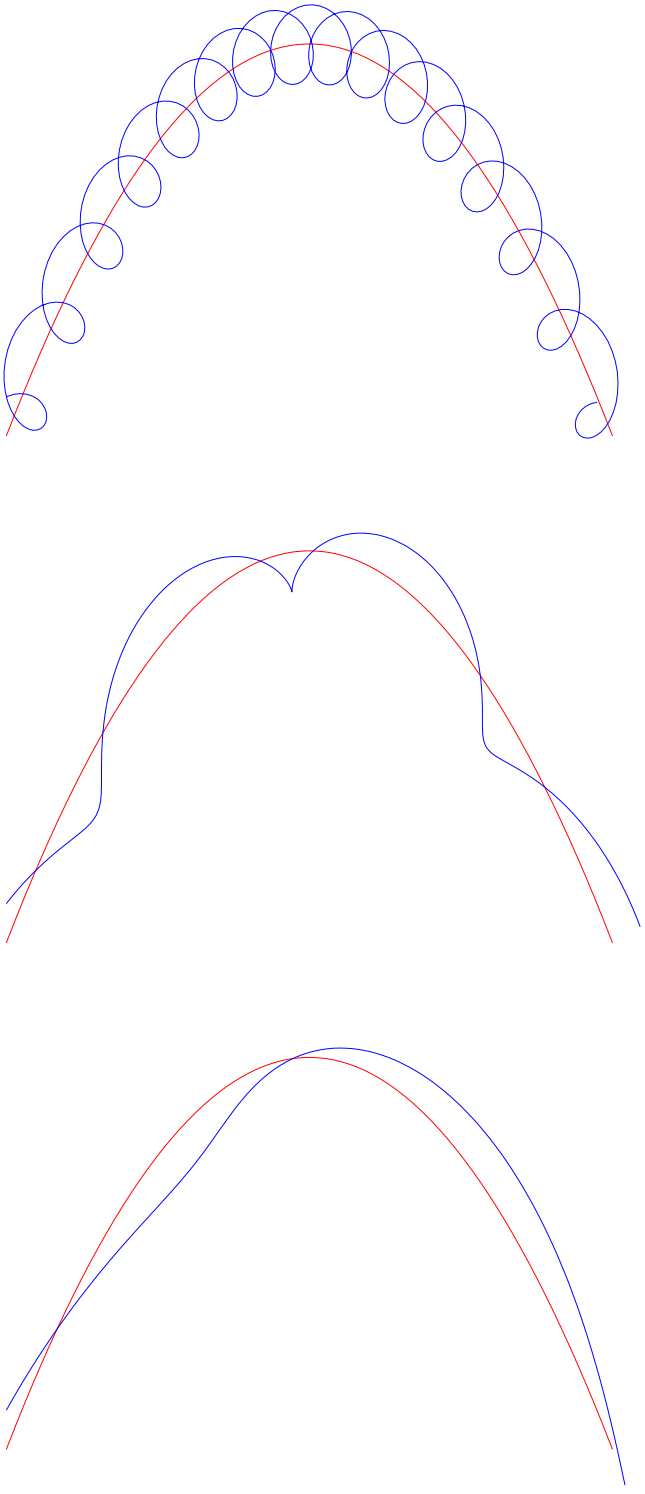


Figure 7.15: The Many Paths to Success

Anyway, as noted above, the total angular momentum of a system is conserved when no net external torque acts on the system. Thus, just as we set up conservation of linear momentum for translating bodies by

$$\sum_{i=1}^{n} \mathbf{p}_{i,before} = \sum_{i=1}^{n} \mathbf{p}_{i,after}$$
$$\sum_{i=1}^{n} m_i \mathbf{v}_{i,before} = \sum_{i=1}^{n} m_i \mathbf{v}_{i,after}$$

we can set up conservation of angular momentum for rotating bodies by

$$\sum_{i=1}^{n} \mathbf{L}_{i,before} = \sum_{i=1}^{n} \mathbf{L}_{i,after}$$

The complication is that whereas $\mathbf{p} = m\mathbf{v}$ is the only expression for linear momentum, there are two expressions for angular momentum:

$$\mathbf{L} = I\omega \qquad \text{or} \qquad \mathbf{L} = \mathbf{r} \times \mathbf{p}$$

The first of these applies when the body is undergoing a pure rotation about an axis, the second when a body is translating. It might seem odd that a body has angular momentum even when it is moving along a straight line, but that in fact turns out to be the case, and when we set up our relations for conservation of angular momentum we must therefore be careful to include contributions of both sorts.

Consider a pirouetting figure-skater: as you know from watching figure skating,[20] when the skater's arms are held out, the rotation is relatively slow; when they are drawn in, the rotation is relatively fast. In this case, the skater is undergoing a pure rotation about an axis, so that the expression we want to use for angular momentum is $I\omega$. When the skater's arms are drawn in, mass is being moved closer to the axis of rotation, so that the moment of inertia I becomes smaller. Since the value of $I\omega$ has to remain constant, this means that the angular velocity ω increases to compensate.[21]

Now consider a pendulum with a changeable length, which you can easily construct in your room or at home with some sort of mass (such as a set of keys) and a bit of string or the like. Drape the string over a finger and set the mass swinging, as shown in fig. (7.16). For a swinging point mass like this, it turns out that we could equally well express the angular momentum as either

[20] It is hard to understand why people watch figure skating. The best we can figure is that, just as the only reason people watch car races is to see the crashes, the only reason people watch skating competitions is for the spectacular falls and wrecks. The only difference would be that at figure-skating competitions they don't sell as much in the way of beer or Slim Jims.

[21] Remind me and we'll get out the giant turntable and actually do this.

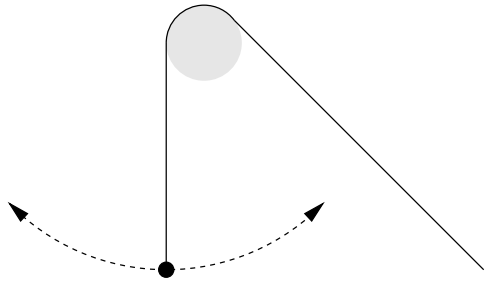


Figure 7.16: A Very Slow Saturday Night

$I\omega$ or $\mathbf{r} \times \mathbf{p}$, the latter of which will reduce to $r_\perp p$ in this two-dimensional example. We'll use the $r_\perp p = r_\perp mv$ here. If, while the mass is swinging, you suddenly pull the string to shorten the length of the pendulum, you are making the $r_\perp$ smaller. Although the pull that you are applying is an external force, it does not exert any torque because it acts along the radial direction — in $\mathbf{r} \times \mathbf{F}$, $\mathbf{r}$ and $\mathbf{F}$ are both along the radial direction centered at the pivot point — , and angular momentum is therefore conserved. To keep the angular momentum $r_\perp mv$ constant, the velocity v at which the mass is swinging must therefore increase. Depending on which way the mass is swinging when you pull the string, you can make it wrap itself around your finger or fly up over and off of your finger. Maybe not the coolest thing you've ever seen, but we needed a simple example of conservation of angular momentum involving $r_\perp p$, and this actually does have practical applications — as those of you know who've gone fishing and gotten your best lure slung over a tree limb on a really bad cast. Not that that ever happened to us.

As a more quantitative example, we return to the merry-go-round: suppose we set a merry-go-round (which we will for simplicity take to be a uniform disk) of radius R and mass m rotating at constant angular velocity ω_0. We now take a standard test cat of mass m_c, dip it in pine sap, and drop it straight down onto the merry-go-round at a distance a from the center, where it sticks like glue. Since there are no external forces with tangential components to exert any torque on the merry-go-round–cat system, angular momentum will be conserved:

$$L_{mgr,i} + L_{c,i} = L_{total,f}$$

The merry-go-round was initially undergoing a pure rotation, so its initial angular momentum will be given by $I\omega$, with I being $\frac{1}{2}mr^2$ for a uniform disk:

$$L_{mgr,i} = \tfrac{1}{2}mR^2\omega_0$$

Had we lobbed the cat onto the merry-go-round like a grenade, then the cat would, depending on the direction of the lob, have had some nonzero angular momentum of the form $\mathbf{r} \times \mathbf{p}$ initially. But since we are simply dropping the

cat straight down onto the merry-go-round, initially it has no motion about the axis of rotation, and so

$$L_{c,i} = 0$$

In the final state we have a compound object consisting of a uniform disk and what amounts to a mass point m_c a distance a from the axis of rotation. The moment of inertia of this compound object is thus

$$I = \tfrac{1}{2}mR^2 + m_c a^2$$

The final angular momentum is therefore

$$L_{total,f} = (\tfrac{1}{2}mR^2 + m_c a^2)\omega_f$$

Our relation for conservation of angular momentum thus becomes

$$\tfrac{1}{2}mR^2\omega_0 = (\tfrac{1}{2}mR^2 + m_c a^2)\omega_f$$

which yields

$$\omega_f = \frac{\tfrac{1}{2}mR^2}{\tfrac{1}{2}mR^2 + m_c a^2}\,\omega_0$$

As you can see, the extent to which the rotational inertia of the cat slows down the rotation of the merry-go-round depends not only on the cat's mass m_c but also on where the cat is placed on the merry-go-round: the greater the distance a from the center the cat is placed, the greater the cat's rotational inertia $I_c = m_b a^2$ and therefore the slower the final angular velocity.

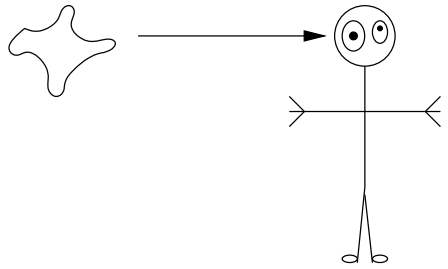


Figure 7.17: ¡Toma ya, Titi!

As another real-life example, suppose you build an antigravity machine and, after setting your roommate floating in midair, plaster him or her upside the head with a huge wad of wet toilet paper, as shown in fig. (7.17). For simplicity, we will take your roommate to be a uniform thin rod of mass m and length ℓ and the toilet paper to be a mass point of mass $\frac{1}{2}m$.[22] The toilet paper strikes your roommate horizontally at speed v_0 and sticks.

[22]Pretty heavy for toilet paper, but maybe its special kryptonite toilet paper that you soaked in mercury or something. We're just trying to keep the algebra simple.

Since there are no external forces or torques acting on this system, both linear and angular momentum will be conserved. Conservation of linear momentum gives, if we denote quantities after the collision by primes,

$$\begin{aligned} p_{tp,0} + p_{rm,0} &= p' \\ (\tfrac{1}{2}m)v_0 + 0 &= (\tfrac{1}{2}m + m)v' \\ v' &= \tfrac{1}{3}v_0 \end{aligned}$$

Conserving angular momentum, however, requires a little thought: About what point should we evaluate our angular momenta? We expect that your roommate, with the toilet paper stuck to his or her head, will tumble clockwise after the collision, and this rotation will be about the center of mass of the compound object consisting of your roommate and the toilet paper. So we must first figure out where this center of mass is located. Measuring from the head end of your roommate, we have

$$x_{cm} = \frac{1}{\frac{1}{2}m + m}\left(\tfrac{1}{2}m(0) + m(\tfrac{1}{2}\ell)\right) = \tfrac{1}{3}\ell$$

Initially, your roommate is stationary and therefore has no angular momentum. About the center of mass of the roommate-toilet paper system (the point that will be center of rotation after the collision), the initial angular momentum of the toilet paper is

$$L_{tp,0} = r_\perp p_{tp,0} = (\tfrac{1}{3}\ell)(\tfrac{1}{2}m)v_0 = \tfrac{1}{6}mv_0\ell$$

After the collision, we have a pure rotation of a compound object consisting of a uniform rod of length $\frac{1}{3}\ell$ and mass $\frac{1}{3}m$ rotated about its end, another uniform rod of length $\frac{2}{3}\ell$ and mass $\frac{2}{3}m$ rotated about its end, and a point mass $\frac{1}{2}m$ a distance $\frac{1}{3}\ell$ from the axis of rotation. The total moment of inertia of this object is thus

$$I = \tfrac{1}{3}(\tfrac{1}{3}m)(\tfrac{1}{3}\ell)^2 + \tfrac{1}{3}(\tfrac{2}{3}m)(\tfrac{2}{3}\ell)^2 + (\tfrac{1}{2}m)(\tfrac{1}{3}\ell)^2 = \tfrac{1}{6}m\ell^2$$

Conservation of angular momentum thus yields

$$\begin{aligned} L_{tp,0} + L_{rm,0} &= L' \\ \tfrac{1}{6}mv_0\ell + 0 &= I\omega' \\ &= \tfrac{1}{6}m\ell^2\omega' \\ \omega' &= \frac{mv_0}{\ell} \end{aligned}$$

Just as it is critically important to get the $\pm$ signs for the forward and backward directions correct when setting up conservation of one-dimensional

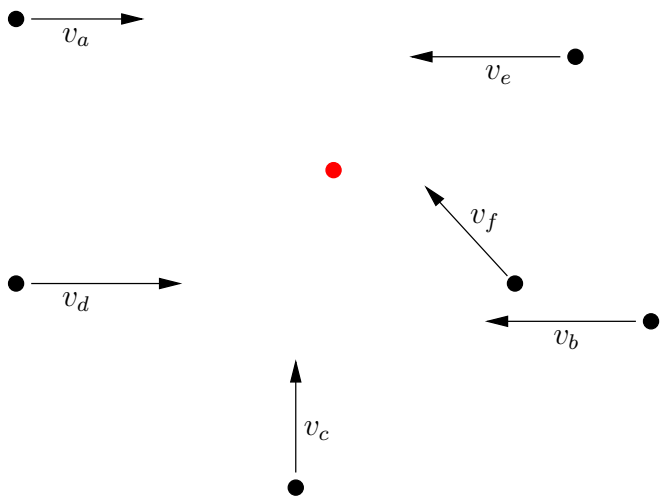


Figure 7.18: Dude!

linear momentum, it is critically important when setting up conservation of angular momentum to get the $\pm$ signs for the clockwise and counterclockwise rotational directions correct. Fig. (7.18) shows the velocity vectors for various contributions of the sort $r_\perp p = r_\perp mv$ to the angular momentum of a system, the center of mass of which is indicated by the red dot. The contributions of v_a, v_b, and v_c to the angular momentum about the center of mass are clockwise; those of v_d, v_e, and v_f are counterclockwise.

In the above roommate-toilet paper example we tacitly took clockwise to be our positive rotational direction. The contribution of the toilet paper to the angular momentum being, like that of v_a in fig. (7.18), clockwise, we wanted it to be positive. Since m, v_0, and ℓ are all positive quantities, $+\frac{1}{6}mv_0\ell$ is positive and was indeed the correct expression for the toilet paper's angular momentum. At the other end of the calculation, our result for ω' was also positive, corresponding to the collision resulting in a clockwise rotation of the system, as we would have expected.

If, however, the toilet paper instead came initially from the right rather than from the left side (like v_e in fig. (7.18)), or if it came from the left and struck *below* the center of mass of the system (like v_d in fig. (7.18)), its contribution to the angular momentum would be counterclockwise, and we should therefore instead set $L_{tp,0} = -\frac{1}{6}mv_0\ell$.

7.8 Kinetic Energy

When a body is simply translating without rotating, the kinetic energy K is $\frac{1}{2}mv^2$. When a body is undergoing a pure rotation, the expression for the kinetic energy is instead $\frac{1}{2}I\omega^2$. But not infrequently, as was the case in the

last example in the preceding section and as is the case for rolling objects, a body is simultaneously translating and rotating. For bodies that are both translating and rotating, you should include both types of contribution to the kinetic energy, so that K is $\frac{1}{2}mv^2+\frac{1}{2}I\omega^2$. These two contributions can in fact be thought of as the rotational version of the separation of kinetic energy into center-of-mass and relative contributions, $K = K_{cm} + K_{rel}$ (eq. (6.13a)), with $K_{cm} = \frac{1}{2}mv^2$ corresponding to the translational motion of the center of mass and $K_{rel} = \frac{1}{2}I\omega^2$ to the relative motion around the center of mass.

In the preceding example, since the toilet paper was initially undergoing a pure translation and your roommate was initially at rest (neither translating nor rotating), the initial total kinetic energy would be

$$K_{tp,0} + K_{rm,0} = \tfrac{1}{2}m_{tp}v_{tp,0}^2 + 0 = \tfrac{1}{2}(\tfrac{1}{2}m)v_0^2 = \tfrac{1}{4}mv_0^2$$

In the final state the roommate-toilet paper system was both translating and rotating, so even though there is only a single "object" after the collision, there are both translational and rotational contributions to the final kinetic energy:

$$\begin{aligned} K' &= \tfrac{1}{2}m_{system}{v'}^2_{system} + \tfrac{1}{2}I_{system}{\omega'}^2_{system} \\ &= \tfrac{1}{2}(\tfrac{3}{2}m)(\tfrac{1}{3}v_0)^2 + \tfrac{1}{2}\left(\tfrac{1}{6}m\ell^2\right)\left(\frac{mv_0}{\ell}\right)^2 \\ &= \tfrac{1}{6}mv_0^2 \end{aligned}$$

The loss of kinetic energy in this inelastic collision would thus be

$$\Delta K = \tfrac{1}{6}mv_0^2 - \tfrac{1}{4}mv_0^2 = -\tfrac{1}{12}mv_0^2$$

This might seem to violate our previous assertion that when two objects stick together the collision is completely inelastic and all the relative kinetic energy $\frac{1}{2}\mu v_{rel}^2$ is lost: here

$$\mu = \frac{m(\frac{1}{2}m)}{m+\frac{1}{2}m} = \tfrac{1}{3}m$$

and the relative velocity, since your roommate was initially at rest, is just the velocity v_0 of the toilet paper, so that

$$K_{rel} = \tfrac{1}{2}\left(\tfrac{1}{3}m\right)v_0^2 = \tfrac{1}{6}mv_0^2$$

which is greater than the $\frac{1}{12}mv_0^2$ actually lost in the collision. But for a collision to be completely inelastic *all* of the relative motion must be lost, and to the extent that there is a rotation about the center of mass afterward, there is still relative motion in the final state. To make the collision completely

inelastic, the toilet paper would have had to hit your roommate dead center, which (as you can immediately see from the symmetry) would result in an absence of rotational motion after the collision.

When two-body collisions include rotational motion, it is in fact quite possible to lose *more* than $\frac{1}{2}\mu v_{rel}^2$ of kinetic energy. As a simple example, consider two identical disks, rotating at equal angular speeds in opposite directions along a common axle, that lock rigidly together after being brought together infinitesimally slowly. In such a case, there is no translational motion before the collision, so that $\frac{1}{2}\mu v_{rel}^2 = 0$. There is, however, still a loss of kinetic energy: since the disks differ only in their directions of rotation, their total angular momentum, and hence their common angular velocity after the collision, is zero, so that all of the initial rotational kinetic energy is lost.

The moral of all this is simply that when collisions include rotational motion, there are also rotational contributions to the relative kinetic energy, and in order for the loss of kinetic energy to be maximal and the collision to be considered completely inelastic the rotational as well as translational contributions to the relative kinetic energy must be killed.

7.9 Torque Due to Gravity

For the special case of gravity near the Earth's surface, the net torque on a system due to the gravitational forces on it is, by eq. (7.42) on p.382,

$$\boldsymbol{\tau} = \sum_{i=1}^{n} \mathbf{r}_i \times m_i \mathbf{g}$$

where the m_i are the bits of mass of which the system is composed. Using the definition (6.2) of the location $\mathbf{r}_{cm}$ the system's center of mass from p.292, we can gerrymander this net gravitational torque into a simple form:

$$\begin{aligned}
\boldsymbol{\tau} &= \sum_{i=1}^{n} \mathbf{r}_i \times m_i \mathbf{g} \\
&= \left(\sum_{i=1}^{n} m_i \mathbf{r}_i \right) \times \mathbf{g} \\
&= m \left(\frac{1}{m} \sum_{i=1}^{n} m_i \mathbf{r}_i \right) \times \mathbf{g} \\
&= m \mathbf{r}_{cm} \times \mathbf{g} \\
&= \mathbf{r}_{cm} \times m \mathbf{g}
\end{aligned}$$

where m is the total mass of the system. This is exactly the result we would have for the torque if all of the system's mass were collapsed to a point at its center of mass. Thus we have the important result that

*The force m***g** *of gravity effectively acts at the center of mass of a system.*

For this reason, the center of mass is sometimes also called the center of gravity. This result, however, depends on the acceleration due to gravity being the same for all the various parts of the system and does not hold for the more general Newtonian inverse-square law of gravity,

$$F = \frac{Gm_1m_2}{r^2}$$

Under this more general law, gravity in general does not act effectively at the center of mass, and in fact it turns out not even to be possible to define a center of gravity in a useful way: while at any given moment it is still possible to determine to single point at which gravity effectively acts, this point changes as the system moves. The only exception to this are spherically symmetric distributions of mass, which are always gravitationally equivalent to point masses.

7.10 Rolling

When you skip along face-first into a tree, your motion is purely translational. When you pirouette, your motion is purely rotational. When you pirouette as you skip along face-first into a tree, your motion is a complicated combination of translational and rotational. Analyzing simultaneous translations and rotations can sometimes be quite involved, but you should definitely have a firm handle on a commonly occurring special case for which the translational and the rotational motions turn out to be related in a simple way: rolling without slipping.

Consider the case of a wheel rolling along level ground. If there is no slipping, then the arc length through which the wheel rotates is equal to the distance that the wheel moves forward, as shown in fig. (7.19). Thus,

$$s = \theta r$$

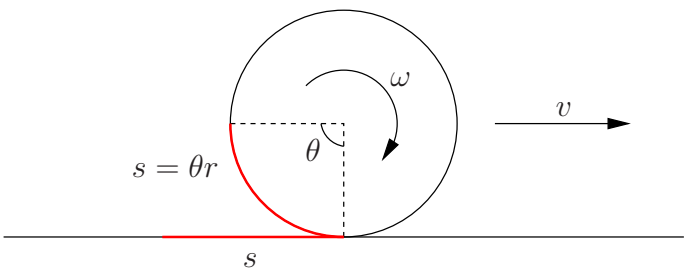


Figure 7.19: Death Awaits You All with Nasty Big Sharp Teeth!

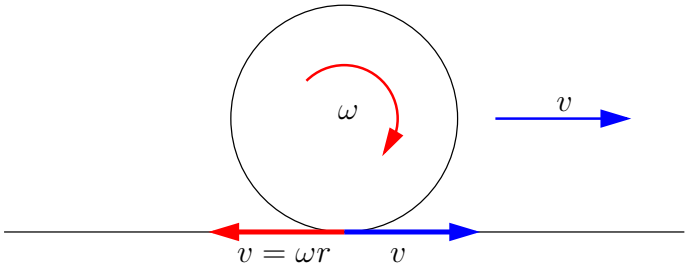


Figure 7.20: Just Look at the Bones!

where θ is the angle through which the wheel has rotated, r is the wheel's radius, and s, because there is no slipping, is both the arc length through which the wheel has rotated and the distance the wheel as a whole has moved forward along the ground. Taking the time derivative of $s = \theta r$ yields

$$v = \omega r$$

where ω is the wheel's angular rate of rotation and v is *both* the tangential velocity that the points on the perimeter of the wheel have by virtue of the wheel's rotation *and* the forward velocity of the wheel as a whole. Taking another time derivative yields

$$a = \alpha r$$

where α is the wheel's angular acceleration and a is both the tangential acceleration experienced by points on the perimeter of the wheel by virtue of changes in the wheel's rate of rotation and the forward acceleration of the wheel as a whole.

Thus the forward translational motion of the wheel and the rotational motion of the wheel about its axis are related by the three relations of table (7.1) on p.356:

$$s = \theta r \qquad v = \omega r \qquad a = \alpha r$$

This is not a trivial statement: usually the s, v, and a in these relations refer only to the tangential motion of points on a rotating body; here we are saying that the s, v, and a also refer to the forward translational motion of the body. One immediate and general consequence of this is that the point of contact (that is, the point on the perimeter of the wheel that is in contact with the surface along which the wheel is rolling) is always stationary, because at that point the forward velocity v shared by all points on the wheel is exactly canceled by the backward tangential velocity $v = \omega r$ due to the wheel's rotation, as shown in fig. (7.20). This lack of motion between the wheel and the surface at the contact point corresponds to there being no slipping during the rolling.

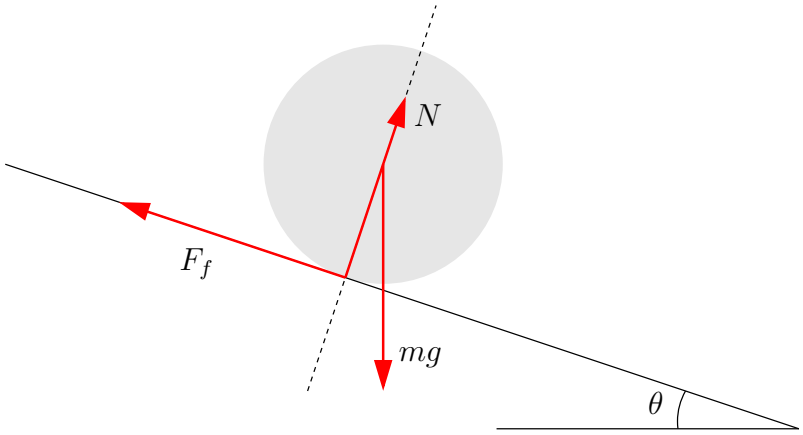


Figure 7.21: A Rolling Cylinder

As an example, consider a uniform solid cylinder of mass m and radius R (and hence moment of inertia $I = \frac{1}{2}mR^2$) rolling down an incline at angle θ to the horizontal. The force diagram for the cylinder is shown in fig. (7.21). Note that the forces are all drawn at the points where they act: the normal N and friction F_f from the point of contact between the cylinder and the incline, and the weight mg of the cylinder from its center. Note also that there must be at least some friction, otherwise the cylinder wouldn't roll — it would just slide down the incline.

There are two basic ways to analyze the motion of the cylinder: by forces and torques, or by conservation of energy. We will first analyze the motion by forces and torques.

For the translational motion, we set up $\mathbf{F} = m\mathbf{a}$. For two-dimensional motion along an incline, this means we get two relations, one for the components parallel to the incline and one for the components perpendicular to it. If we make out of the incline the positive direction for the perpendicular force relation, we have

$$N - mg\cos\theta = ma_\perp = 0 \tag{7.47a}$$

where we have noted that since there is no motion in the perpendicular direction, $a_\perp = 0$. And if we make down the incline the positive direction for the parallel force relation, we have

$$mg\sin\theta - F_f = ma \tag{7.47b}$$

For the rotational motion, we set up $\tau = I\alpha$. Each of the three forces acting on the cylinder generates a torque of the form $r_\perp F$. To evaluate these torques, we need to draw the line of each force so that we can drop a perpendicular to it from the axis of rotation (which runs through the center

of the cylinder) and figure out the $r_\perp$. Since the weight mg effectively acts right at the center of the cylinder, its $r_\perp = 0$. And although the normal force N does not act at the center of the cylinder, its line of force passes through the center, so that its $r_\perp = 0$ as well. So the only force that contributes to the net torque on the cylinder is friction. Since friction acts along the incline, its $r_\perp$ is just the radius R of the cylinder. The torque due to friction is thus $r_\perp F_f = RF_f$, and we can see from fig. (7.21) that, about the center of the cylinder, this torque due to friction will be clockwise. The issue is now whether we should make clockwise or counterclockwise the positive rotational direction. We are of course free to choose either way, but for reasons that will become clear shortly, the convenient choice is to make clockwise positive. Our rotational relation is thus

$$RF_f = I\alpha = \tfrac{1}{2}mR^2\alpha \tag{7.47c}$$

The mass m and radius R of the cylinder and the angle θ of the incline are known; the unknowns in eqq. (7.47a), (7.47b) and (7.47c) are the frictional force F_f, the normal force N, and the angular and linear accelerations α and a of the cylinder. Eq. (7.47a) can be solved at once for the normal force N:

$$N = mg\cos\theta$$

This leaves us with two equations ((7.47b) and (7.47c)) in three unknowns (F_f, α, and a). In order to solve for these unknowns, we need another relation among them. It is here that the condition of rolling without slipping comes in: we also have $a = \alpha R$. Together, $a = \alpha R$ and eqq. (7.47b) and (7.47c) give us three equations in three unknowns.

This is also where we see that choosing to make clockwise the positive rotational direction was convenient: we made down the incline the positive direction for the translational motion, and the cylinder will roll clockwise when rolling down the incline. By making clockwise the positive rotational direction, we have been consistent about directions between the translational and rotational motions, so that $a = \alpha R$. Had we chosen to make counterclockwise the positive rotational direction, then we would have had $a = -\alpha R$ — equally correct and quite viable, but arguably not as convenient. From this newly enlightened point of view, our reasoning back when we were choosing the signs for the rotational directions would be as follows: we already chose down the incline to be the positive direction for the translational motion, and when the cylinder rolls down the incline it rotates clockwise, so we should choose clockwise as the positive rotational direction.

When solved, the system of $a = \alpha R$ and eqq. (7.47b) and (7.47c) yields

$$a = \tfrac{2}{3}g\sin\theta \qquad \alpha = \tfrac{2}{3}\frac{g}{R}\sin\theta \qquad F_f = \tfrac{1}{3}mg\sin\theta$$

There are several things to observe about these solutions:

- Recall that for a mass sliding without friction down an incline, the acceleration is $g \sin\theta$. The acceleration of the rolling cylinder is somewhat less than $g \sin\theta$ because the friction required for the rolling retards the motion.

- Looked at another (somewhat looser) way, for the rolling cylinder, the same pull mg of gravity as in the case of sliding has to provide not only the translational motion down the incline but also the rotational motion. As a result, less of the mg goes into the translational motion, which means that the acceleration down the incline is slower.

- The value of the frictional force is determined entirely by the condition that the cylinder is rolling without slipping. We did not use (and in fact it would have been incorrect to use) a relation like $F_f = \mu N$ to determine the frictional force.

  Since the points on the cylinder that are in contact with the incline are always stationary, it is static friction that applies, and static friction always adjusts itself to whatever value is necessary to prevent motion, from zero up to some maximal value. Our solutions are telling us that the amount of friction necessary for rolling without slipping is $F_f = \frac{1}{3} mg \sin\theta$. If this is less than the maximal value of static friction, the static frictional force will adjust itself to this value in order to prevent motion at the point of contact. If $\frac{1}{3} mg \sin\theta$ exceeds the maximal force static friction can generate, then there will be slipping and our analysis above is not valid (starting from the point where we set $a = \alpha R$).

  Again,

  > *The value of the frictional force is determined entirely by the condition that the cylinder is rolling without slipping. In general,* $F_f \neq \mu_s N$.

  We cannot emphasize this enough — as we know from bitter experience. We could dress up like a giant chicken and scream this repeatedly through a megaphone while dancing a tarantella, and any number of you would still prove perversely intent on using $F_f = \mu_s N$. Perhaps it's simply a reflex that has become so hardwired into the autonomic nervous system that it is beyond conscious control, or perhaps its a malicious conspiracy to convince us of the utter futility of existence. But even in the latter case, beware of unintended consequences: if for no other reason, out of enlightened self-interest you shouldn't want to us have a total nervous breakdown in front of the class and go postal. That would be just ugly. So remember: The value of the frictional

force is determined entirely by the condition that the cylinder is rolling without slipping. Your future may depend on it.

- The coefficients $\frac{1}{3}$ and $\frac{2}{3}$ depended on the moment of inertia of the cylinder, which was $\frac{1}{2}mR^2$. Objects with different moments of inertia will have different coefficients in the solutions. For a cylindrical shell, for example, the moment of inertia would have been mR^2 and the acceleration and frictional force would have worked out to $a = \frac{1}{2}g\sin\theta$ and $F_f = \frac{1}{2}mg\sin\theta$: the hoop has a larger moment of inertia, so it requires more frictional torque to set it rotating, and consequently its motion down the incline is slower.

Now let's analyze the motion of the cylinder by conservation of energy. Because the cylinder is both translating and rotating, there will be both the rotational $\frac{1}{2}I\omega^2$ and the translational $\frac{1}{2}mv^2$ contributions to the kinetic energy. The relevant potential energy is the gravitational mgy. Although there is a frictional force, no energy is lost to friction and there is therefore no need to modify our relation for conservation of energy to take this friction into account: because the point of contact between the cylinder and the incline is stationary, the corresponding displacement, and hence the work done by friction, are zero. Conservation of energy for the cylinder thus takes the form

$$\begin{aligned} K_i + U_i &= K_f + U_f \\ \tfrac{1}{2}I\omega_i^2 + \tfrac{1}{2}mv_i^2 + mgy_i &= \tfrac{1}{2}I\omega_f^2 + \tfrac{1}{2}mv_f^2 + mgy_f \\ \tfrac{1}{2}(\tfrac{1}{2}mR^2)\omega_i^2 + \tfrac{1}{2}mv_i^2 + mgy_i &= \tfrac{1}{2}(\tfrac{1}{2}mR^2)\omega_f^2 + \tfrac{1}{2}mv_f^2 + mgy_f \end{aligned}$$

The mass of the cylinder cancels out, leaving us with

$$\tfrac{1}{4}R^2\omega_i^2 + \tfrac{1}{2}v_i^2 + gy_i = \tfrac{1}{4}R^2\omega_f^2 + \tfrac{1}{2}v_f^2 + gy_f \tag{7.48}$$

Since the cylinder is rolling without slipping, we have $v = \omega R$ at all times in the motion. How we use $v = \omega R$ in conjunction with eq. (7.48) depends on the situation.

Suppose we know that the cylinder starts from rest and we want to figure out the cylinder's speed at the after it has rolled a distance ℓ down the incline. Since the cylinder starts from rest, $\omega_i = v_i = 0$. Also, a distance ℓ along the incline corresponds to a vertical drop of $\ell\sin\theta$, so we can set $y_i = \ell\sin\theta$ and $y_f = 0$. With these substitutions, eq. (7.48) becomes

$$0 + 0 + g(\ell\sin\theta) = \tfrac{1}{4}R^2\omega_f^2 + \tfrac{1}{2}v_f^2 + 0$$

Since we want to solve for v_f, we would now use $v = \omega R$ in the form $\omega_f = v_f/R$ to eliminate ω_f:

$$g\ell\sin\theta = \tfrac{1}{4}R^2\left(\frac{v_f}{R}\right)^2 + \tfrac{1}{2}v_f^2 = \tfrac{3}{4}v_f^2$$

which yields

$$v_f = \sqrt{\tfrac{4}{3} g\ell \sin\theta}$$

Note that this is the same solution we would obtain by using the acceleration $a = \frac{2}{3} g \sin\theta$:

$$\begin{aligned} v^2 - v_0^2 &= 2a(x - x_0) \\ v^2 - 0 &= 2(\tfrac{2}{3} g \sin\theta)(\ell) \\ v &= \sqrt{\tfrac{4}{3} g\ell \sin\theta} \end{aligned}$$

7.10.1 All Good Things Must Come to an End

Our results of the preceding section indicate that when the object is rolling on level ground ($\theta = 0$) we have, no matter what the value of the moment of inertia, $F_f = 0$, $a = 0$, and $\alpha = 0$: when an object is rolling along level ground, there is no net force or torque on it, and its velocity and angular velocity both remain constant (with $v = \omega r$). Under these ideal circumstances, the object would simply continue rolling forever at the same linear and angular speeds. But this leads to the question, if there is no frictional force or torque, how then, as we observe in the real world, do objects rolling on level surfaces slow down and come to rest?

It might seem that we have only to have a nonzero backward frictional force between the rolling object and the surface, but that wouldn't work out: while such a backward frictional force would cause the linear motion of the object to slow down, it would give rise to a torque that would actually speed up the rotation.

The dynamics are a bit involved, but what actually happens is basically this: the velocity and angular velocity would be constant only in the ideal case that both the object and the surface were perfectly smooth and hard. Real objects and surfaces are a bit rough and "mushy," at least on a microscopic scale, so that we have one of the three situations shown (with greatly exaggerated scale) in fig. (7.22): the object meets with small obstructions on the surface, sinks a bit into the surface, or is deformed a bit by the surface. The forces exerted on the object that have a backward horizontal compo-

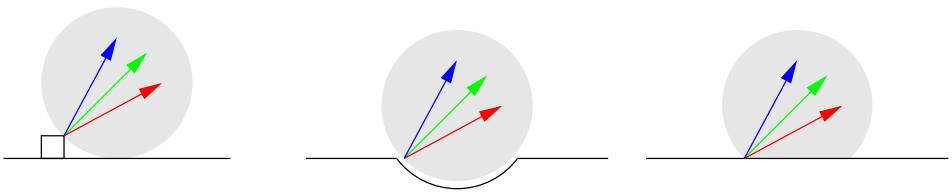


Figure 7.22: So It Goes

nent, and that therefore tend to slow down the linear motion of the object, then fall into one of three cases: those that pass through the center of the object and therefore contribute zero torque (the green arrows in fig. (7.22)), and those that pass to one side or the other of the center, giving rise to forward torques that would speed up the rotation (the red arrows) or backward torques that would slow down the rotation (the blue arrows). There may be a combination of all three types of forces acting, but if there is no slipping on a macroscopic scale, the forces with backward torques predominate and slow the rotation of the object down in sync with its linear motion, so that v and ω both decrease but always $v = \omega r$.

7.11 Massive Pulleys

Massive pulleys don't sound very exciting, and they aren't. You have only to keep in mind that when a pulley has a mass or there is friction in its bearings, the tension will in general differ on the two sides of the pulley.

As an example, consider the situation shown in fig. (7.23), in which the (frictionless) pulley is a uniform disk of mass m_p and radius R, the blocks have masses m_1 and m_2, and there is no friction on the horizontal surface. Setting up $F = ma$ for the translational motion of the two blocks gives

$$\begin{aligned} T_1 &= m_1 a \\ m_2 g - T_2 &= m_2 a \end{aligned}$$

where we have been careful to be consistent about our signs: when m_1 moves to the right, m_2 moves down. Setting up $\tau = I\alpha$ for the pulley gives

$$\begin{aligned} \tau &= I\alpha \\ RT_2 - RT_1 &= \tfrac{1}{2} m_p R^2 \alpha \end{aligned}$$

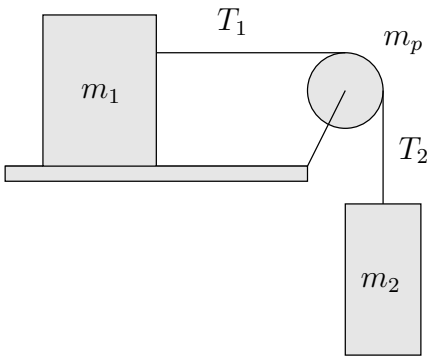


Figure 7.23: Prelude to a Snooze

where we have again been careful to be consistent about our signs: when m_1 moves to the right, the pulley rotates clockwise, and the downward pull of T_2 and leftward pull of T_1 tend to rotate the pulley clockwise and counterclockwise, respectively. Although two other forces act on the pulley in addition to the tensions — the pulley's own weight and a supporting force, of unknown magnitude and direction, exerted on the pulley by its axle — neither of these other two forces exerts any torque on the pulley because they both act (at least effectively, in the case of the pulley's weight) at the pulley's center.

If there is no slipping of the rope over the pulley, we have $a = \alpha R$, which constitutes the fourth equation needed to solve for the four unknowns T_1, T_2, a, and α. The solutions for these unknowns are

$$a = \frac{m_2}{m_1 + m_2 + \frac{1}{2}m_p}\, g \qquad \alpha = \frac{m_2}{m_1 + m_2 + \frac{1}{2}m_p}\, \frac{g}{R}$$

$$T_1 = \frac{m_1 m_2 g}{m_1 + m_2 + \frac{1}{2}m_p} \qquad T_2 = \frac{(m_1 + \frac{1}{2}m_p)m_2 g}{m_1 + m_2 + \frac{1}{2}m_p}$$

Nothing very surprising here. As you can see from the $\frac{1}{2}m_p$ in the denominators of the expressions for a and α, the pulley's rotational inertia retards the angular and hence linear accelerations.

7.12 Gyroscopes & Tops

A fuller quantitative treatment of gyroscopes and tops would require far more time and math than we have at our disposal — general gyroscopic motion is notoriously difficult to tackle — , but there are some basics and qualitative features you should be aware of.

The relation (7.40)

$$\boldsymbol{\tau}_{net\ ext,\ system} = \frac{d\mathbf{L}_{total,\ system}}{dt}$$

told us that in the absence of a net external torque a system's angular momentum is conserved. In this relation the torque and the angular momentum are both vectors, and conserving angular momentum as a vector means keeping not only its magnitude but also its direction constant, so that neither the rate at which the body is rotating nor the direction of its axis of rotation change.

A gyroscope is basically just a spinning object. Those used for navigation are housed in a system of bearings and gimbals designed to minimize any torque exerted on the gyroscope, so that the orientation of its axis of rotation can serve as a fixed reference. In practice, there is always some friction in the bearings and gimbals. One effect of this friction is to slow the gyroscope's

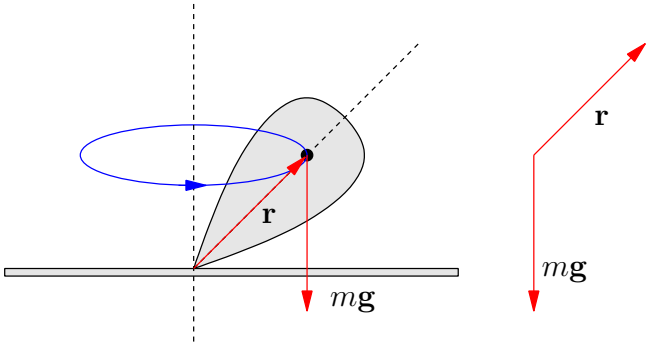


Figure 7.24: Precession of a Spinning Top

rotation down. To prevent this, gyroscopes are built with electric motors to maintain their angular velocity. Frictional torque also tends to change the direction of the gyroscope's axis of rotation. To prevent this, gyroscopes are by design very heavy (to give them large moments of inertia I) and kept rotating at very high angular velocities $\boldsymbol{\omega}$, so that they have very large angular momenta $\mathbf{L} = I\boldsymbol{\omega}$. The change $d\mathbf{L}$ in angular momentum due to frictional torque is thus made proportionally smaller.

Fig. (7.24) shows a top spinning on a table. The top's angular momentum will, as a vector, be along its axis of rotation. About the (fixed) point of contact with the table, there is a torque due to gravity:

$$\boldsymbol{\tau}_{grav} = \mathbf{r} \times m\mathbf{g} \tag{7.49a}$$

From fig. (7.24), you can see that $\mathbf{r} \times m\mathbf{g}$ will be into the page ($\otimes$). Since

$$\boldsymbol{\tau} = \frac{d\mathbf{L}}{dt} \tag{7.49b}$$

this means that the change $d\mathbf{L}$ in the top's angular momentum $\mathbf{L}$ will be into the page and thus perpendicular to $\mathbf{L}$ itself. Recall that when the change in a vector is perpendicular to the vector, what is changed is the vector's direction; its magnitude remains constant. For the top, this means that its axis of rotation will revolve around the vertical, as indicated by the blue ellipse in fig. (7.24). This effect is familiar to everyone and is called the precession.

If the top is already precessing according to eqq. (7.49), then that is the end of the story: the top simply continues to precess around the vertical at a constant rate and angle of tilt. Tops are, however, more usually released from initial states that do not satisfy eqq. (7.49), and an analysis of the motion in this more general case is very involved. It turns out that the top will settle pretty much into the precession of eqq. (7.49), but as it does so it will develop wobbles called nutations. Depending on the values of the various parameters

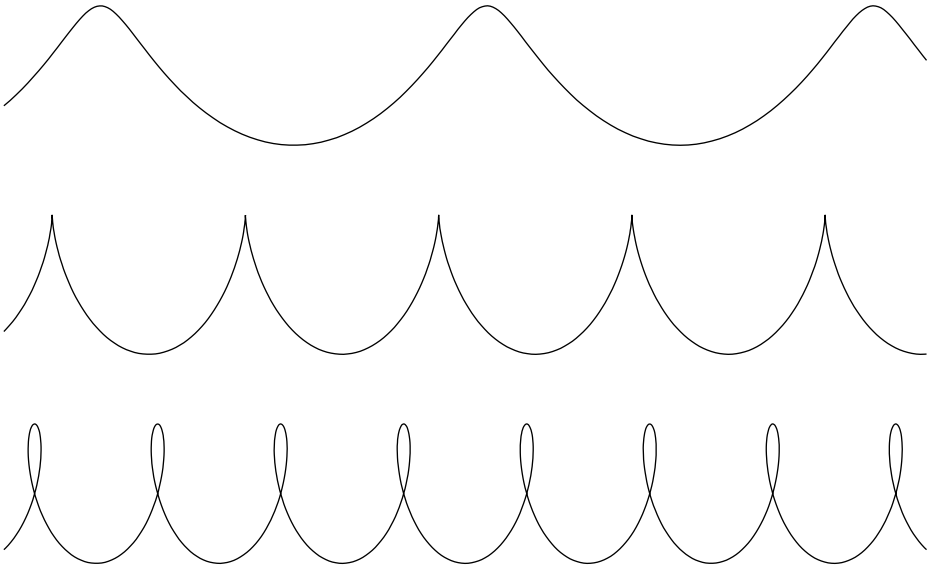


Figure 7.25: Nutations of a Spinning Top

(the top's rate of spin, etc.), these nutations can look like any of the three cases in fig. (7.25): wavy, cuspy, or loopy. These nutations are often small and may have escaped your notice unless you've watched closely.

7.13 Summary of Important Points

- For every translational quantity and relation, there is an analogous rotational quantity and relation, as given in table (7.1) on p.356 (with the full vector forms derived in §7.2). Among these are the relations for constant angular acceleration (§7.4 on p.373), which you should be able to remember and apply.

  The rotational equivalent of mass is the moment of inertia,

  $$I = \sum_{i=1}^{n} m_i r_{\perp,i}^2 \qquad \text{or} \qquad \int dm\, r_\perp^2$$

  Table (7.2) on p.374 gives the moments of inertia for some common symmetric distributions of mass (which you should be able to work out, if need be). You should also be able to apply the parallel-axis theorem (§7.6 on p.379).

  The rotational equivalents of momentum and force are angular momentum and torque. The angular momentum of a body moving with momentum $\mathbf{p}$ is

  $$\mathbf{L} = \mathbf{r} \times \mathbf{p}$$

The torque due to a force $\mathbf{F}$ is

$$\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}$$

In two dimensions, these relations reduce to

$$L = r_\perp p \qquad\qquad \tau = r_\perp F = r F_{tan}$$

with $\pm$ signs for clockwise and counterclockwise. For a body undergoing a pure rotation about an axis, the angular momentum and torque are

$$L = I\omega \qquad\qquad \tau = I\alpha$$

Torque and angular momentum are related by

$$\boldsymbol{\tau} = \frac{d\mathbf{L}}{dt}$$

- The values you get for angular momenta and torques depend on the points about which they are evaluated. Relations among angular momenta and torques are of course not meaningful unless you have been consistent and evaluated all of those angular momenta and torques about the same point.

  When you draw force diagrams for bodies undergoing rotational motion, for the purpose of evaluating torques you should draw all the force vectors from the points at which those forces act on the body. Remember that the gravitational force mg acts effectively at the body's center of mass.

- Unless a body is constrained to rotate about some other point, it will naturally rotate about its center of mass. In particular, a body on which there is no net external force will rotate about its center of mass.

- In situations involving both rotating and translating bodies, you want to set up $\mathbf{F} = m\mathbf{a}$ for the translational motion and $\tau = I\alpha$ for the rotational motion.

  For example, for suspended blocks or blocks on surfaces connected by cables slung over massive pulleys, you will want to set up $\mathbf{F} = m\mathbf{a}$ for each of the blocks and $\tau = I\alpha$ for each of the pulleys. For rolling objects, you set up both $\mathbf{F} = m\mathbf{a}$ and $\tau = I\alpha$.

  When doing an energy calculation, remember that the kinetic energy is $\frac{1}{2}mv^2$ for translating bodies, $\frac{1}{2}I\omega^2$ for rotating bodies, and $\frac{1}{2}mv^2 + \frac{1}{2}I\omega^2$ for bodies that are both translating and rotating.

  Note that even though the pivot is not at the center of mass, an object swinging about a pivot, like a pendulum, is undergoing a pure rotation. You take into account that the pivot is not at the center of mass by using the moment of inertia *about the pivot*, not about the center of mass.

- Since a cable always exerts its tension tangent to the pulley, the $r_\perp$ used to calculate the torque due to the tension is always simply the radius of the pulley.

  In general there is no relation between the rotational and translational motions of a body; they are independent. If, however, the cable passes over the pulley without slipping, then $s = \theta r$, $v = \omega r$, and $a = \alpha r$. You can use the latter to relate the translational acceleration of blocks connected to the cable to the angular acceleration of the pulley and thus relate the a in the $F = ma$ for the block to the α in the $\tau = I\alpha$ for the pulley. When you do this, be careful to be consistent about the signs for direction, so that the positive direction for the translational motion of the block matches the positive direction (clockwise or counterclockwise) for the rotational motion of the pulley.

- Again, while in general the rotational and translational motions of a body are independent, if a body is rolling without slipping, then $s = \theta r$, $v = \omega r$, and $a = \alpha r$. You can use the latter to relate the body's translational acceleration to the angular acceleration of its rotation in $\mathbf{F} = m\mathbf{a}$ and $\tau = I\alpha$. When you do this, be careful to be consistent about the signs for direction, so that the positive direction for the translational motion of the body matches the positive direction (clockwise or counterclockwise) for its rotational motion.

  Unless you are dealing with a really bizarre situation, it is friction that supplies the torque needed for the rotational motion of the rolling object. How much friction is needed for the rolling ultimately depends on the body's moment of inertia relative to its mass and will be determined by $\tau = I\alpha$. *Do not assume that this frictional force obeys* $F_f = \mu N$. In the system of three equations for the body — $F = ma$, $\tau = I\alpha$, and $a = \alpha r$ — the frictional force is frequently one of the three unknowns you are solving for.

- When there is no net external torque on a system, its total angular momentum is conserved.

  When you conserve angular momentum in collisions, be mindful that not only do rotating bodies make contributions of the form $I\omega$, but translating bodies make contributions of the form $\mathbf{r} \times \mathbf{p}$ or $r_\perp p$. If a body is translating as it simultaneously rotates about an axis, its angular momentum will have contributions of both forms.

  For rotating bodies, or bodies simultaneously translating and rotating, it is convenient to use eq. (7.45):

$$\mathbf{L} = \mathbf{L}_{of\ cm} + \mathbf{L}_{about\ cm}$$

  The $L_{of\ cm}$ will be of the form $r_\perp p$ (from any translational motion the body has); the $L_{about\ cm}$ will be of the form $I\omega$.

7.14 Problems

1. In a stunning display of animal psychosis, your dog, starting from rest, takes off barking and running in circles of radius r.

   (a) If your dog accelerates at a constant rate and completes two revolutions as the time goes from $t = 0$ to $t = T$,

      i. What is your dog's angular acceleration?

      ii. How fast is he moving at time T?

      iii. How fast will he be moving when he completes his third revolution?

      iv. How long will it take him to complete n revolutions?

   (b) If instead his constant acceleration gets him up to speed v when he has completed his first two complete revolutions,

      i. What is his angular acceleration?

      ii. How long did it take him to complete these first two revolutions?

      iii. How fast will he be moving when he completes his third revolution?

2. Recall that in problem #1 in Chapter 3 you had escaped at bath time and were running around outside, shrieking with joy, in your birthday suit. Now one of your parental units chases you in a circle of radius r. At $t = 0$, you and your parental unit are diametrically opposite each other (that is, 180° apart), with you running away at a constant angular velocity ω_0 and your parental unit pursuing at $\frac{1}{2}\omega_0$. Your parental unit turns on the afterburners and generates enough angular acceleration to catch you after you have completed one more complete revolution. What was your parental unit's angular acceleration?

3. The orbital radii of Mars and Earth about the Sun are $r_m = 2.279 \times 10^{11}$ m and $r_e = 1.49597870660 \times 10^{11}$ m, respectively. How long between syzygies, that is, between occasions when the Earth, Mars, and the Sun lie along a straight line? The masses of the Mars, Earth, and the Sun are $m_m = 6.4185 \times 10^{23}$ kg, $m_e = 5.9723 \times 10^{24}$ kg, and $m_s = 1.98844 \times 10^{30}$ kg, respectively. See the footnote if you need a hint.[23]

[23] Think in angular terms.

4. Fan blades look like just a blur to us when they are rotating at full speed, but everyone must have noticed that as the blades are speeding up when the fan is turned on (or slowing down when it is turned off) they appear to alternate between rotating forward and rotating backward. This happens because our vision works more or less like a movie camera, taking not a continuous picture but rather a series of still photos. If the effective "frame rate" (frames per second, or fps) of your vision is f and the fan has four blades,

   (a) For what ranges of values of ω will the blades appear to be rotating forward? See the footnote if you need a hint.[24]

   (b) For what ranges of values of ω will the blades appear to be rotating backward?

   (If you remind me, we can get a fan and a strobe light and play around with this.)

5. Some utterly silly person designs a door where the knob is mounted in the middle of the door. Why is this a really stupid design?

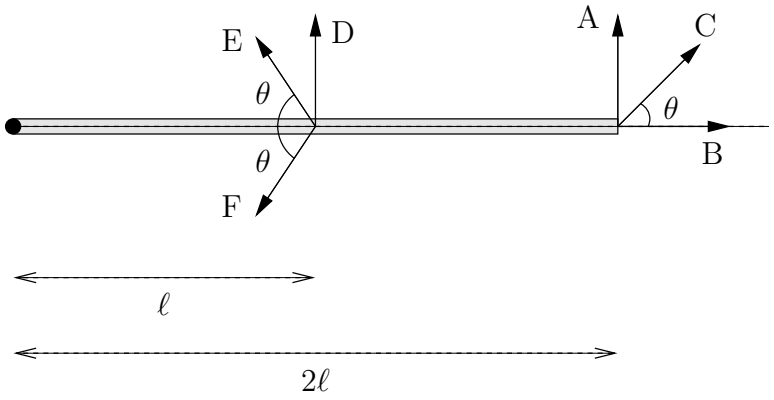


Figure 7.26: Problem 6

6. The ludicrously poor diagram (7.26) is supposed to depict the handle of a wrench. The dot at the left end represents the bolt you are trying to loosen, and the vectors all represent a force F_{you} that you are applying at various places and angles.

   (a) Determine the torque you exert in each case.

   (b) What is the difference in the torques for cases E and F?

[24] Think about the position of the blades in the next frame relative to their position in the previous frame.

7. For reasons known only to you, you try to set a flywheel rotating by cranking it by hand. The flywheel is a solid disk of mass of m and radius r, and you are careful always to exert a constant force F_{you} at a right angle to the crank-arm, which is of length ℓ.[25]

   (a) What torque are you applying to the flywheel?

   (b) What is the angular acceleration of the flywheel?

   (c) How much energy will you have put into the flywheel after you have cranked for a time t?

   (d) How much power are you putting into cranking the flywheel, as a function of this time t?

8. A uniform thin rod of length ℓ and mass m is rotated about a perpendicular axis through a point one quarter of the way from one of its ends. Show that the moment of inertia is $\frac{7}{48}m\ell^2$.

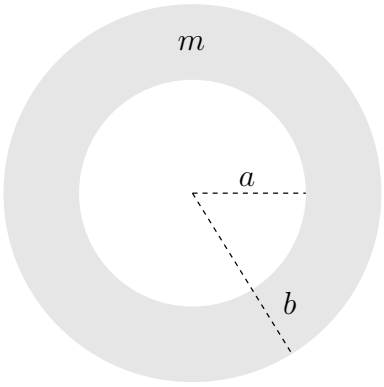


Figure 7.27: Problem 9

9. Fig. (7.27) shows an annulus[26] of mass m and inner and outer radii a and b, respectively. The mass m is uniformly distributed over the annulus.

   (a) Show that the moment of inertia of the annulus about its axis (that is, the axis through its center and perpendicular to its plane) is

   $$I = \tfrac{1}{2}m(a^2 + b^2)$$

   (b) Show that you can reproduce the moments of inertia of a solid disk and a thin ring from this result.

[25] This is of course very unrealistic; whether you use your arms or legs, you would tend rather to pump the crank, in a way similar to that by which the piston rod rotates the wheel of a steam locomotive. The same physiological limitation affects the pedaling of a bicycle.

[26] Recall that *annulus* is just a snobby word for a washer.

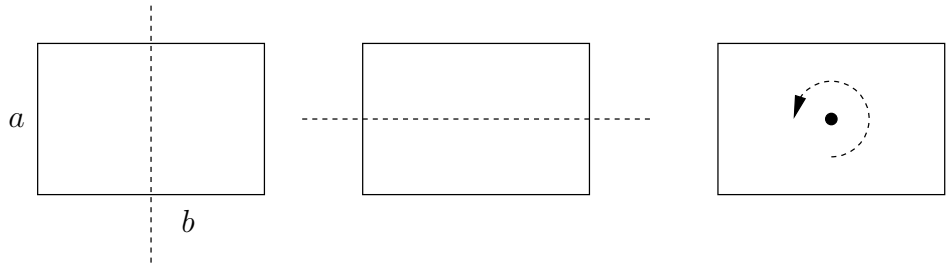


Figure 7.28: Problem 10

10. On a really slow Saturday night, you construct a rectangular frame out of wire of uniform linear density λ and twirl it about various axes, as shown in fig. (7.28). The lengths of the sides of the rectangle are a and b.

    (a) Determine the moment of inertia of the rectangle about an axis
        i. Through the center of the rectangle, as shown on the left side of fig. (7.28).
        ii. Through the center of the rectangle, as shown in the middle of fig. (7.28).
        iii. Through the center of the rectangle and perpendicular to its plane, as shown on the right side of fig. (7.28). See the footnote if you need a hint.[27]

    (b) Show that your results for #10(a)i through #10(a)iii reduce to what you expect in the special case $a = 0$ (or, equivalently, $b = 0$).

    (c) You should find that your answers to #10(a)i and #10(a)ii add up to your answer for #10(a)iii. By setting up Cartesian coordinate axes, show that for any planar distribution of mass in the xy plane

    $$I_{xx} + I_{yy} = I_{zz}$$

    where I_{xx} is the moment of inertia about the x axis, etc. See the footnote if you need a hint.[28]

    (d) Is your proof in the preceding part limited to cases where the axes pass through the center of mass of the system?

    (e) Repeat this proof, starting from eq. (7.19),

    $$I = \int dm \begin{pmatrix} r^2 - x^2 & -xy & -xz \\ -xy & r^2 - y^2 & -yz \\ -xz & -yz & r^2 - z^2 \end{pmatrix}$$

    for rotations about the x, y, and z axes. See the footnote if you need a hint.[29]

[27] Use the parallel-axis theorem.

[28] When you apply the definition of the moment of inertia about each axis, express $r_\perp^2$ in terms of x, y, and z.

[29] Express r^2 in terms of x, y, and z, and note that rotations about the coordinate axes

11. A turntable (essentially a solid disk of mass m_t and radius a) is rotating at angular frequency ω_0. A 12 inch LP record (mass m_r, also essentially a solid disk of radius a) is gently dropped onto it.

    (a) (History question.) At how many revolutions per minute do turntables for 12 inch LPs rotate?

    (b) After some ugly slipping and skidding, the record and the turntable will rotate together at the same angular speed. What makes this happen?

    (c) Under the assumption that there is no net external torque on the system, what is the final angular speed of the record and turntable?

    (d) Was the assumption that there is no net external torque on the system valid?

    (e) Is this collision elastic or inelastic? If inelastic, what is the change in kinetic energy, and how do you account for this change physically?

12. (a) The original wheel, as designed by Oogg the Cave Man, was a solid disk. Consider a solid wheel of mass m and radius r rolling along at speed v. What is the wheel's total kinetic energy, and what fraction of this is rotational?

    (b) As an improvement on this design, Oogg's wife, Ooggina, invents the spoke. With spokes, a wheel of the same radius has mass m' and becomes essentially a thin ring. (That is, the mass of the spokes is small enough that it can be neglected, so that effectively all the mass is in the perimeter of the wheel.) At the same speed v, what is the wheel's total kinetic energy, and what fraction of this is rotational?

    (c) What then is the advantage of a spoked over a solid wheel?

13. You are "sleeping" your yo-yo, which we will treat as essentially a uniform solid disk of radius r. When you wake it up with a very slight nudge, you find that it climbs only a vertical distance ℓ back up toward your (stationary) hand.

    (a) Show that while it was sleeping the yo-yo was rotating at angular frequency
    $$\omega = \frac{2\sqrt{g\ell}}{r}$$

    (b) Why did you not need to take into account work done by the tension while the yo-yo was climbing?

---

correspond to
$$\vec{\omega} = \begin{pmatrix}\omega\\0\\0\end{pmatrix} \quad \text{or} \quad \begin{pmatrix}0\\\omega\\0\end{pmatrix} \quad \text{or} \quad \begin{pmatrix}0\\0\\\omega\end{pmatrix}$$

14. A uniform solid disk and a thin ring, both of mass m and radius r, are rolled at the same speed v along a level surface leading into an incline at angle θ to the horizontal. If there is no slipping, which rolls farther up the incline? Prove your assertion.

15. A model American family is pleasantly picnicking in a peaceful meadow at the Plucking Pheasant Farm when, without warning, a huge freaking boulder of bone-crushing mass m rolls from rest down the mountainside and squashes them all. The boulder is essentially a uniform sphere of radius R, and to reach the family it rolls a distance ℓ down an incline at angle θ to the horizontal.

    (a) Determine how fast the boulder is moving when it mows down the family.

    (b) Determine the boulder's acceleration down the incline.

    (c) Determine the frictional force acting on the boulder.

    (d) For what physical reason do your results for the preceding parts depend or not depend on the mass or radius of the boulder? You may be strongly tempted merely to cite the obvious algebraic reasons and get on with your life, but you should think about the physical effects underlying what happens algebraically.

    (e) What implications would this have for a ball of yarn that was unraveling as it rolled down an incline?

    (f) Based on #15d, what would you expect to see in an avalanche in which more or less round boulders of various sizes and densities roll down a mountainside?

    (g) If Hollywood uses fake boulders consisting of papier mâché shells in an avalanche scene, how would the motion of the fake boulders differ from the real thing? Of course these days Hollywood cheats and does everything by digital special effects, but for the purposes of this problem let's pretend we are back in the good old days when they used to use miniature models and other physical constructions.

16. Some round object of mass m and radius r rolls without slipping down an incline at θ to the horizontal. *Yawn.* Though we suppose you could make it the Round Object of Rolling Death and put an orphanage at the bottom of the incline or something like that to spice things up. Anyway, the object's acceleration down the incline is a constant $\frac{1}{3}g\sin\theta$.

    (a) Determine the moment of inertia of the object and the frictional force between the object and the incline.

    (b) Is this possible? That is, could there be an object that would experience an acceleration of $\frac{1}{3}g\sin\theta$ while rolling without slipping down an incline at angle θ to the horizontal?

17. Why don't you have to take into account the work done by the frictional force when applying conservation of energy to objects rolling down inclines?

18. Determine the frictional force needed to keep an object rolling without slipping at a constant speed along a level surface.

19. We have seen that the frictional force required for a uniform solid cylinder of mass m and radius r to roll without slipping down an incline at angle θ to the horizontal is $\frac{1}{3}mg\sin\theta$. Suppose, however, that there is insufficient friction for rolling without slipping and that as the cylinder skids from rest down the incline a constant frictional force of only $\frac{1}{6}mg\sin\theta$ acts on it.

    (a) Determine the linear and angular accelerations of the cylinder.

    (b) Determine the velocity of the cylinder when it has descended a distance ℓ along the incline.

    (c) How much work is done by friction as the cylinder covers this distance ℓ? If you are inclined to say none, or have arrived at that conclusion by calculation, you will want to check the assumptions you made about things angular. See the footnote if you need a hint.[30]

20. Rework problem #19 for the more general case of a nonuniform cylinder with moment of inertia $I = \beta mr^2$, where $0 \leq \beta \leq 1$, with a frictional force $\kappa mg\sin\theta$.

    What physical restrictions are there on the value of κ and hence on values of the linear and angular accelerations of the cylinder? And what light does this shed on #16?

[30] You need to be mindful that the slipping affects the extent of the rotational motion. It may help to think in terms of the time it takes the cylinder to cover the distance ℓ; in this way you can determine both the extent of the rotational motion and the distance through which the cylinder has effectively slipped.

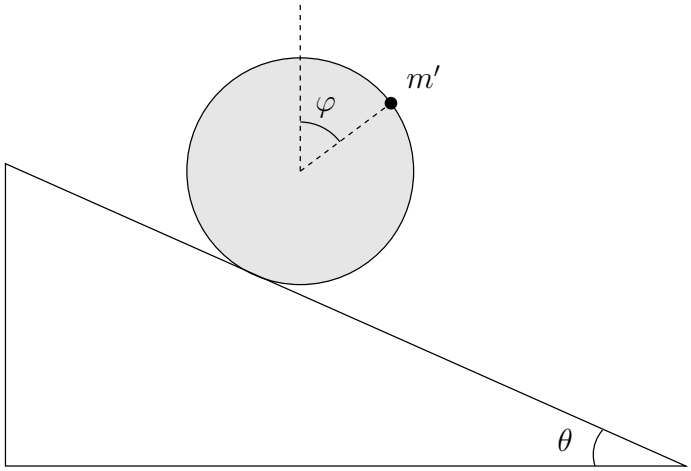


Figure 7.29: Problem 21

21. You make a personal statement by sticking a half-chewed Gummy Bear of mass m' on the perimeter of a uniform disk of mass m and radius R and releasing the disk from rest on an incline tilted at angle θ to the horizontal.

    Show that when, as shown in fig. (7.29), the Gummy Bear is at angle φ to the vertical the translational acceleration of the disk down the incline is

$$a = \tfrac{1}{2}\frac{(m+m')g\sin\theta + m'g\sin\varphi + m'\omega^2 R\sin(\varphi-\theta)}{m+m'\Big(1+\cos(\varphi-\theta)\Big)}$$

    where ω is its angular velocity at that instant.

22. A marble of mass m and radius r rolls around a loop-the-loop of radius $R \gg r$.

    (a) What minimal speed does the marble have to have at the bottom of the track to just make it around the top?

    (b) Now do the calculation for the case that $r \lesssim R$, that is, for the case when the size of the marble is not necessarily negligible compared to the size of the loop-the-loop.

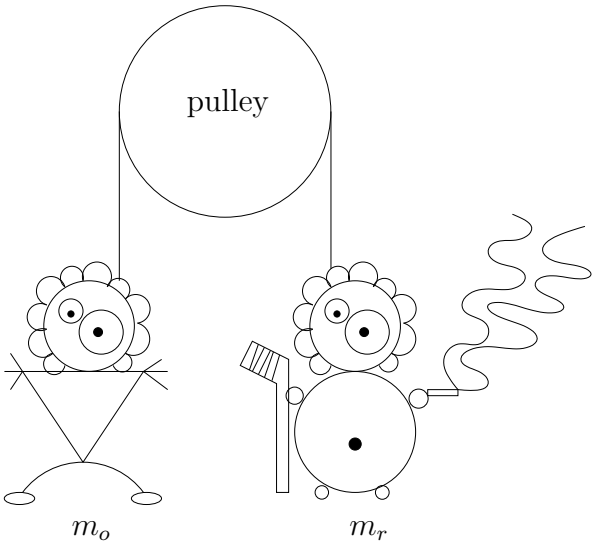


Figure 7.30: Problem 23

23. (The Return of Atwood.) Recall that in our last episode (#23 on p.231), the pulley in fig. (7.30) was massless and the tension was the same throughout the cord connecting the two masses.

    Now the pulley is no longer massless, but rather a uniform solid disk of mass m and radius R.

    (a) Determine the tensions on either side of the pulley, the acceleration of the masses, and the angular acceleration of the pulley. (After you get a result for one tension, you should be able to simply write down the result for the other tension without doing any calculation.)

    (b) Explain how your results for the preceding part make sense when

        i. $m_r \gg m_o$ (or vice versa).

        ii. $m_r = m_o$.

        iii. $m \to 0$.

        iv. m large.

    (c) What force must be exerted on the pulley (by its axle or whatever it is mounted on) to hold it up?

    (d) Now analyze the machine by energy methods and determine the speed of the masses after the heavier mass has descended from rest through a distance h.

    (e) Show that this result agrees with what you get by using your result for the acceleration of the masses.

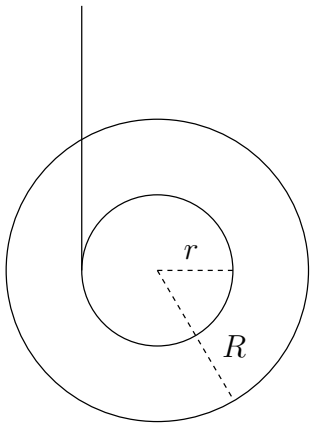


Figure 7.31: Problem 24

24. Fig. (7.31) is a sort of x-ray view of a yo-yo. The upper end of the cord is held at rest. The yo-yo has mass m and moment of inertia I, and its spindle, around which the cord is wrapped, has a radius r that for no really good reason is greatly exaggerated in size in the figure.

    (a) Determine the acceleration of the yo-yo and the tension in the cord.

    (b) How might you increase the moment of inertia of the yo-yo without changing its mass, the radii r and R, or the axis of rotation? Similarly, how might you increase the mass m of the yo-yo without changing its moment of inertia or the radii r and R?

    (c) If you hold m, r, and R constant, what therefore are the maximal and minimal possible values of the acceleration of the yo-yo?

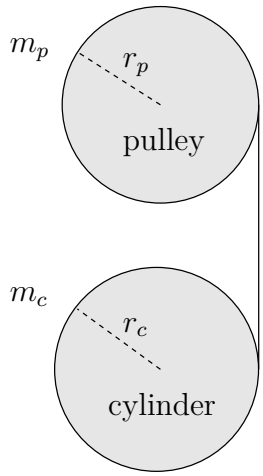


Figure 7.32: Problem 25

25. Fig. (7.32) shows a pulley (essentially a uniform solid disk) of mass m_p and radius r_p and a uniform solid cylinder of mass m_c and radius r_c. One end of a cord is wrapped around the pulley, so that the cord unwinds as the pulley rotates. The other end of the cord is wrapped around the cylinder, so that the cord also unwinds from the cylinder as it falls.

    (a) Determine the acceleration of the cylinder. You will want to be careful about how you relate the various accelerations (the downward acceleration of the cylinder, the angular acceleration of the pulley, and the angular acceleration of the cylinder); see the footnote if you need a hint.[31]

    (b) Make physical sense of your result in the limit $m_p \to 0$.

    (c) You should find that in the limit $m_c \to 0$ your result doesn't make sense. Why is this case is giving you nonsense? See the footnote if you need a hint.[32]

[31] Think about arc lengths and how the amount of cord that unwinds from the pulley and cylinder are related to the distance the cylinder falls.

[32] The reason is mathematical, not physical.

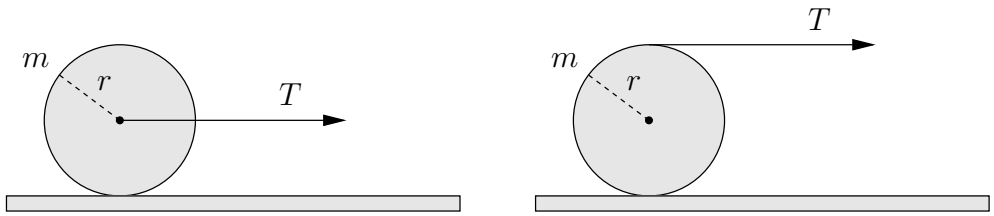


Figure 7.33: Problem 26

26. The left side of fig. (7.33) shows a uniform solid cylinder of mass m and radius r being pulled along a level surface by a tension T applied horizontally to its axle. The right side of fig. (7.33) shows an identical cylinder, also being pulled along a level surface by a horizontally applied tension — but get this: this time the cord is wound around the cylinder, so that the tension is applied at its edge!!! Okay, so maybe that's not really all that exciting. But you can't say we didn't try. AND AT LEAST WE DIDN'T USE ALL CAPS. Anyway, you should treat the axle as massless, which is a polite way of saying that the axle itself does not enter into the problem at all — it's just a device for applying the tension to the center of the cylinder on the left side of fig. (7.33).

    (a) If the surface is frictionless, what are the acceleration and angular acceleration of the cylinder

        i. On the left side of fig. (7.33)?

        ii. On the right side of fig. (7.33)?

    (b) If there is enough friction with the surface that the cylinder is rolling without slipping, what are the acceleration and angular acceleration of the cylinder and the frictional force exerted on the cylinder by the surface

        i. On the left side of fig. (7.33)?

        ii. On the right side of fig. (7.33)?

    (c) Your result for #26(b)ii may seem surprising, but explain how it makes sense physically and show that the rate at which work is being done by the applied tension does in fact equal the rate of change of the kinetic energy of the cylinder. See the footnote if you need a hint.[33]

    (d) In fact, while we're at it, show that the rate at which work is being done by the applied tension equals the rate of change of the kinetic energy of the cylinder in #26(a)i, #26(a)ii, and #26(b)i as well.

[33] Think about how the length of cord that unwinds from the cylinder, and therefore the distance through which the tension acts, is related to the distance through which the cylinder translates and the angle through which it rotates — and hence to its linear and angular accelerations.

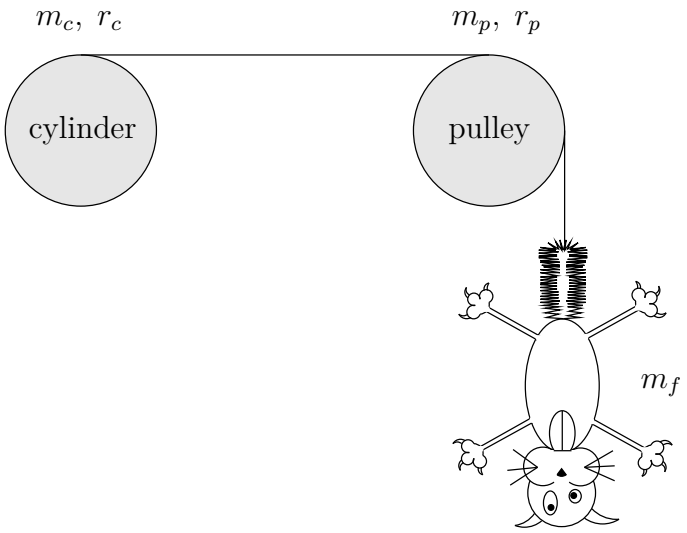


Figure 7.34: Problem 27

27. In a bizarre experiment, you wrap a cord around a cylindrical pulley (a uniform solid cylinder of mass m_c and radius r_c), run it over another pulley (a uniform solid disk of mass m_p and radius r_p), and from the other end suspend a cat of mass m_f (f for *feline*, to avoid confusion with c for *cylinder*).

    (a) Determine the acceleration of the cat.

    (b) What limits of the mass m_f, m_p, and m_c can you take to check your result for the acceleration? Show that you get what you expect in these limits.

    (c) Suppose that instead of being a pulley, the mass m_c can roll (without slipping) over a level surface as the cord unwinds from it. Determine the accelerations of the cylinder and the cat. Be careful of how you relate the acceleration of the cat to the linear and angular accelerations of the cylinder; if you have thought about it and cannot determine this relation, checkout the footnote.[34]

[34] In notation that should be clear, $a_f = a_c + \alpha_c r_c$.

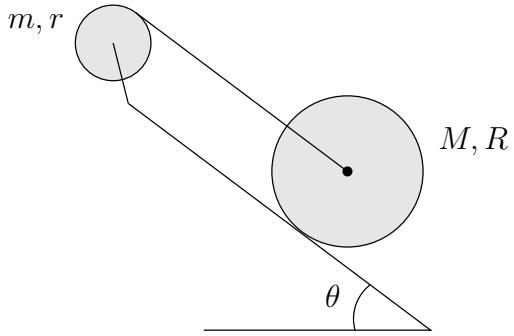


Figure 7.35: Problem 28

28. Believe it or not, fig. (7.35) shows a cord running from a pulley (which we will take to be a uniform solid disk of mass m and radius r) to the axle of a uniform solid cylinder of mass M and radius R that is rolling without slipping down an incline at angle θ to the horizontal. Not really terribly interesting, but maybe you can imagine there's a kitten or a small child at the bottom of the incline. Anyway, the geometry is such that, very conveniently, the cord is parallel to the incline, and you can treat the axle as massless (in other words, the axle itself does not enter into the problem at all — it's just a device for applying the tension to the center of the cylinder).

    (a) Determine the frictional force between the cylinder and the incline, the acceleration of the roller, and the tension in the cord.

    (b) What limits of the masses m and M and of the angle θ can you use to check your result for the acceleration? Show that you get what you expect in these limits.

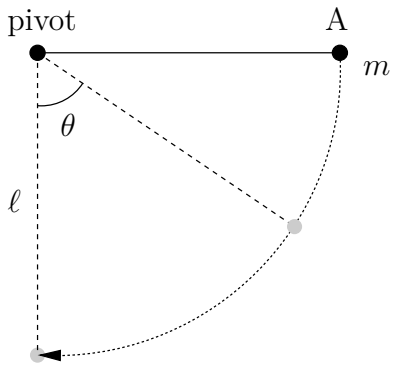


Figure 7.36: Problem 29

29. The technical term simple pendulum refers to the idealized case of a point mass (the "bob") attached to a pivot by means of a massless rod or cord. In fig. (7.36), the length of the pendulum is ℓ, and the bob (mass m) is released from rest at point A, which is level with the pivot.

    (a) Use rotational methods to determine how fast the bob is moving as a function of the angle θ that the pendulum makes with the vertical.

    (b) What is the tension in the cord at the lowest point?

    (c) Determine the tangential and radial components of the acceleration of the bob as a function of θ.

30. A pendulum with a more complicated distribution of mass than the simple pendulum of $^{\#}29$ is called a compound pendulum.

    (a) Suppose that *instead of* a point-mass bob at the end of a massless rod or cord, the pendulum in $^{\#}29$ consists of a uniform rod of mass m pivoted about one end. Determine how fast the far end of the rod is moving as a function of the angle θ that the pendulum makes with the vertical. See the footnote if you need a hint.[35]

    (b) Now suppose that *in addition to* the bob of mass m, the rod of the pendulum in $^{\#}29$ has a mass m'. Determine how fast the bob is moving as a function of θ.

[35] Remembering the importance of the center of mass in the gravitational potential energy will make your life easier. And who among us doesn't wish that life were easier?

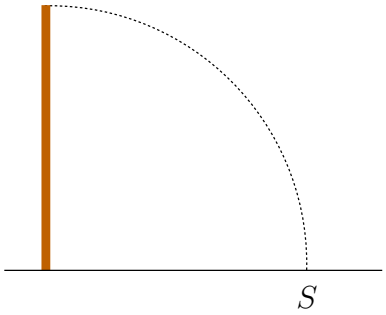


Figure 7.37: Problem 31

31. You chop down a tree, which we will rather unimaginatively regard as a uniform rod of mass m and length ℓ. When the tree pops and starts to fall, you run, in exactly the wrong direction, and not quite far enough, to point S, as shown in fig. (7.37). (Here, by the principle known to psychoanalysts as *over-determination*, S stands for both *stupid* and *splat*.)

    (a) How fast is the tip of the tree moving when it makes you a candidate for a Darwin Award?

    (b) You should have found that your result for #31a was greater than the speed of a body dropped from rest through a height ℓ. Explain how this makes sense physically.

    (c) Determine the radial and tangential components of the acceleration of the tip of tree as a function of the angle θ that it makes with the vertical.

    (d) Determine the vertical component of the acceleration of the tip of tree as a function of the angle θ that it makes with the vertical.

    (e) You should have found that the radial, tangential, and vertical accelerations all exceed $1\,g$ at some point during the fall. In what temporal order do they reach $1\,g$?

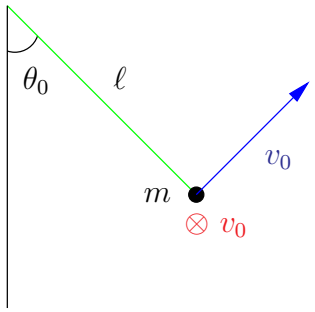


Figure 7.38: Problem 32

32. Fig. (7.38) shows a tetherball: a mass m connected to one end of a massless green cord of length ℓ, the other end of which is connected to a swivel mounted on a black pole, so that the ball is free to swing in any direction over a spherical surface of radius ℓ under the influence of only the tension and gravity. We have used a boring point mass m in fig. (7.38) just to keep the diagram simple, but if you want you can imagine that the "ball" is instead a cat, a baby, a grenade, a bottle of concentrated sulfuric acid, or whatever. Anyway, we will assume that the cord never goes slack, so that the velocity of the ball is always tangential, and we will restrict our attention to two special cases:

    (a) The velocity v_0 of the ball at the instant depicted in fig. (7.38) is in the vertical plane shown by the blue arrow.

    (b) The velocity v_0 of the ball at the instant depicted in fig. (7.38) is in the horizontal plane shown by the red $\otimes$.

    Determine the minimal value of v_0 for which the ball will rise to $\theta = \frac{\pi}{2}$ in each of these cases.

    For the latter case, you will want to think in terms of conservation principles, being very careful to consider for each potentially conserved quantity whether the conditions of conservation are satisfied in this particular situation. See the footnote if you need a further hint.[36]

    You should find, as you would expect, that your solutions to #32a and #32b are degenerate when $\theta_0 = 0$.

[36] You will want to consider each component of the vector angular momentum separately.

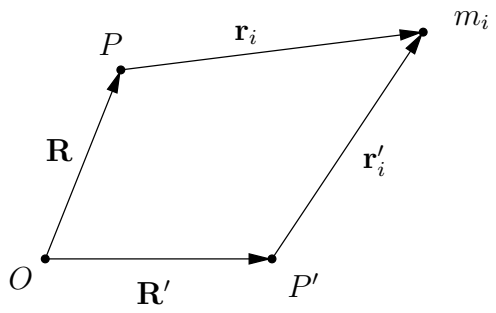


Figure 7.39: Problem 33

33. A system is composed of mass points m_i that have velocities $\mathbf{v}_i$. As shown in fig. (7.39), we have chosen a point O as the origin of our coordinate system, and we will be concerned with the value of the total angular momentum of the system about two points: P and P', the locations of which are given by the position vectors $\mathbf{R}$ and $\mathbf{R}'$, respectively. The location of m_i relative to P and P' is given by $\mathbf{r}_i$ and $\mathbf{r}'_i$, respectively.

    (a) Show that the difference between the value $\mathbf{L}$ of the total angular momentum of the system about point P and its value $\mathbf{L}'$ about point P' is

    $$\mathbf{L} - \mathbf{L}' = (\mathbf{R} - \mathbf{R}') \times \mathbf{P}$$

    where $\mathbf{P}$ is the total linear momentum of the system. See the footnote if you need a hint.[37]

    (b) Interpret this result.

[37] Start by writing down expressions for $\mathbf{L}$ and $\mathbf{L}'$ based on the definition of angular momentum, then think about what relation fig. (7.39) suggests among $\mathbf{R}$, $\mathbf{R}'$, $\mathbf{r}_i$, and $\mathbf{r}'_i$.

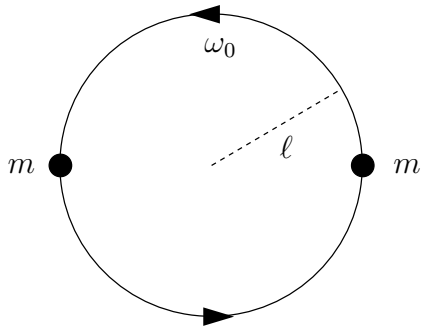


Figure 7.40: Problem 34

34. Let us for simplicity idealize a figure skater, spinning with arms extended, as a pair of point masses m (representing the arms) rotating in a circle of radius ℓ at angular velocity ω_0, as shown from the perspective of someone looking down at the rink from above in fig. (7.40).

    (a) Determine the kinetic energy and angular momentum of this pair of masses.

    (b) The skater now pulls his or her arms (that is, the point masses) in to radius $\kappa\ell$, where $0 < \kappa < 1$. Determine the angular velocity, kinetic energy, and angular momentum of the pair of masses as a function of κ.

    (c) Determine, as a function of κ, the force that the skater must apply to pull each of his or her arms in.

    (d) Calculate the work done by this force when the skater draws his or her arms in from radius ℓ to radius $\kappa\ell$ and confirm that it equals the change in kinetic energy. See the footnote if you need a hint.[38]

[38] Remember that the variable is κ, not ℓ: $dr = d(\kappa\ell) = \ell\, d\kappa$.

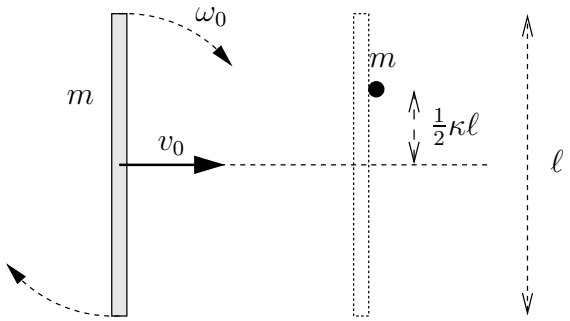


Figure 7.41: Problem 35

35. Fig. (7.41) shows a hockey puck (essentially a point mass) just minding its own business when along comes a hockey stick (which we will for simplicity treat as a uniform thin rod), sliding over frictionless ice at speed v_0 and rotating, *independently*, at angular speed ω_0. The length of the stick is ℓ, and for simplicity we will rather unrealistically make the mass of the stick and the puck both m. The puck is initially offset from the center of the stick by a distance $\frac{1}{2}\kappa\ell$, where $0 \leq \kappa \leq 1$ is a variable parameter, the value of which we will tinker with in later parts of the problem. When the stick collides with the puck, the stick is oriented as shown by the dotted outline in fig. (7.41) and the stick and the puck somehow magically stick together. Maybe we should have thought of some better scenario than a hockey stick and puck and then we wouldn't have had to invoke magic. But it's not like the problem wasn't already absurdly unrealistic, anyway.

    (a) Describe the motion of the stick and the puck after the collision. About what point will the stick-and-puck system rotate?

    (b) Determine the moment of inertia of the system about this point.

    (c) Determine the velocity of the system after the collision.

    (d) Someone argues that conservation of angular momentum in the collision gives, as the relation for the angular velocity ω of the system after the collision,

    $$\tfrac{1}{12}m\ell^2\omega_0 = \tfrac{1}{12}m\ell^2(1 + \tfrac{3}{2}\kappa^2)\omega$$

    Set this person straight.

    (e) Determine the angular velocity of the system after the collision, including its direction.

    (f) To keep things from getting too hairy, we will now set $\omega_0 = 0$.

        i. With $\omega_0 = 0$, what value of κ will maximize the kinetic energy lost in the collision?

        ii. Why, physically, does this value of κ result in the maximal loss of kinetic energy?

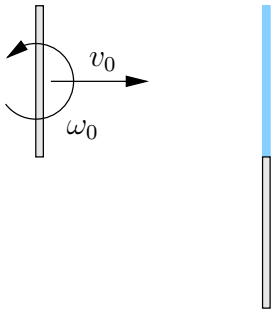


Figure 7.42: Problem 36

36. Far out in the dark and lonely void of space, a uniform thin rod of mass m and length ℓ spins at a counterclockwise angular velocity ω_0 as it translates, *independently*, to the right at speed v_0 in search of a mate, as shown in fig. (7.42). This rod meets up with another rod, identical (that is, also mass m and length ℓ) but initially stationary. At the moment that the rods meet, the moving rod (as indicated by the blue "ghost" in the figure) just happens to be aligned with the stationary rod and the two rods lock rigidly together in dysfunctional co-dependency.

    (a) Describe the motion of the joint rod after the collision.

    (b) Determine

        i. The translational velocity of the joint rod after the collision.

        ii. The angular velocity of the joint rod after the collision.

        iii. The change in the kinetic energy of the system as a result of the collision.

    (c) What would your results for #36(b)i and #36(b)iii be if, instead of between two rods, the collision were between two point masses (and thus, of course, head-on)?

    (d) Make physical sense of the differences between your results for #36c and the limit of your results for #36b as $\omega_0 \to 0$.

    (e) Make physical sense of your results for #36b when $v_0 = 0$.

    (f) Is it possible for the joint rod to have zero angular velocity after the collision? If not, explain the physical reason why not. If so, for what values or ranges of values of the various quantities involved would this come about, and what is happening physically?

    (g) Is it possible for the joint rod to have zero translational velocity after the collision? If not, explain the physical reason why not. If so, for what values or ranges of values of the various quantities involved would this come about, and what is happening physically?

    (h) Is it possible for the collision between rods to be elastic? Prove your assertion.

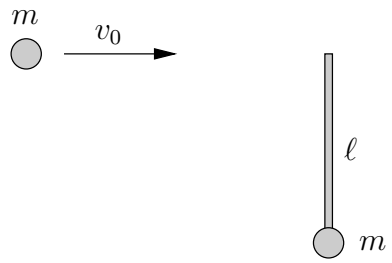


Figure 7.43: Problem 37

37. Out in the void of space, a mass m moving at speed v_0 strikes the end of a stationary rod of length ℓ at a right angle and, wouldn't you know it, sticks right to it. The rod itself is massless, but has another mass, also m, at its other end, as shown in fig. (7.43). (You may treat both of these masses m as point masses. Also, there are no pivot or anchor points; the system is floating freely in empty space.)

    (a) Determine the translational and rotational velocities of the rod after the collision.

    (b) Show that the collision is elastic.

    (c) How do you reconcile the elasticity of the collision with the fact that the masses stick together? See the footnote if you need a hint.[39]

[39] What are the velocities of the two masses immediately after the collision?

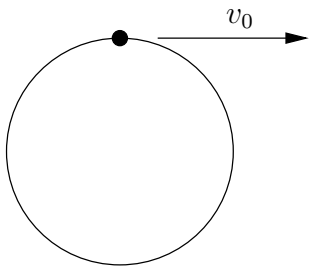


Figure 7.44: Problem 38

38. Out in the void of space, a point mass m is rigidly attached to the perimeter of a uniform ring of radius R and the same mass m. At the instant shown in fig. (7.44) the point mass is at the very top of the ring and moving to the right with velocity v_0, but the center of the ring at that moment is stationary.

    (a) Describe the motion.

    (b) Determine the velocity of the center of mass of the system.

    (c) Determine the acceleration of the center of mass of the system.

    (d) Determine the angular momentum of the system about the center of the ring.

    (e) Determine the angular momentum of the center of mass ($L_{of\ cm}$).

    (f) Determine the angular momentum about the center of mass ($L_{about\ cm}$) about the center of the ring.

    You should find that your results for #38d, #38e, and #38f obey eq. (7.45):

$$L = L_{of\ cm} + L_{about\ cm}$$

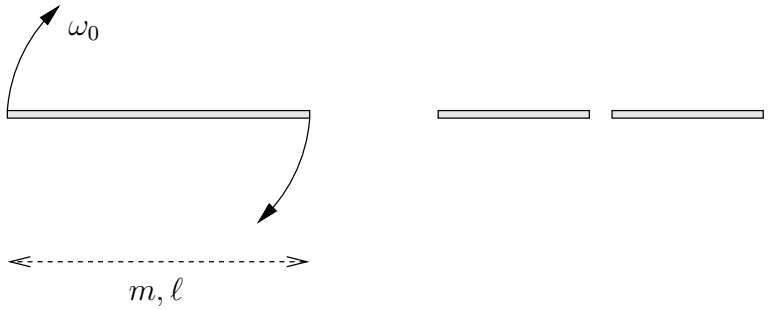


Figure 7.45: Problem 39

39. Recall that in #36 two emotionally needy rods had become rigidly locked together in dysfunctional co-dependence. Such a relationship cannot, of course, last, and now it's time for the divorce:

    A uniform thin rod of mass m and length ℓ is spinning about its center at angular velocity ω_0, without translational motion, as shown on the left side of fig. (7.45), when it suddenly splits into two identical halves.[40] The instant of separation is almost shown on the right side of fig. (7.45) — *almost* meaning that for clarity the figure shows some initial separation between the two halves; in fact, there is no gap between the two halves at the instant of separation. Nor is any energy lost or released when the rod splits.

    Anyway, as the divorce-court judge, you have to determine the rotational and translational velocities of the halves after separation.

    You may find it helpful to think about

    - What symmetry tells you about the final rotational and translational velocities.
    - What quantities are conserved.
    - The direction of the velocity of the various points along each of the two half-rods immediately before the split and what this tells you about the direction of motion of each of the two half-rods immediately after the split.
    - The "average" velocity of the various points along each of the two half-rods immediately before the split and what this tells you about the velocities of the centers of mass of the two half-rods immediately after the split.

    Or not. Whatever.

[40] How it could possibly split into two nonidentical halves we leave for the reader to ponder.

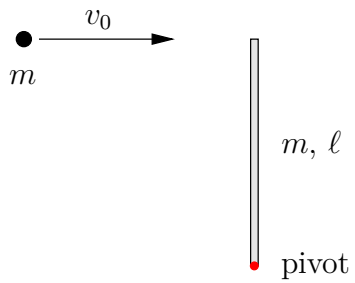


Figure 7.46: Problem 40

40. Far out in the void of space, a rather boring point mass m moving at speed v_0 perpendicularly strikes and sticks to the end of an equally boring uniform thin rod of length ℓ and, conveniently, equal mass m, as shown in fig. (7.46) But get this: the other end of the rod is anchored at a pivot (the red dot in fig. (7.46)).[41] *Bwahahahaha!*

    (a) Determine the resulting angular velocity of the rod.

    (b) Determine the kinetic energy lost in the collision.

    (c) If there were no pivot, so that the rod were completely free to move, what would the velocity of the center of mass of the system be after the collision?

    (d) With the pivot, what is the velocity of the center of mass of the system after the collision?

    (e) Determine the *direction* of the force exerted on the rod by the pivot during the collision (which you can treat as essentially an instantaneous process).

    (f) Make physical sense of this direction.

    (g) Why doesn't this force invalidate your result for #40a?

[41] Okay, so you might ask, Dude, how you can have a pivot point out in the void of space? What's it anchored to? And you would have an excellent point there. Busted. But let's just pretend you could somehow nail that end of the rod to the spacetime continuum and get on with our lives.

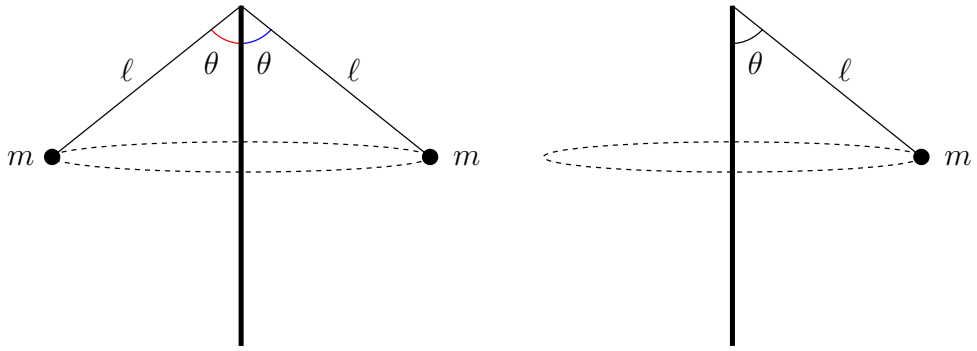


Figure 7.47: Problem 41

41. (The return of the governor, from #36 on p.239.) The left side of fig. (7.47) shows a somewhat simplified version of a mechanical arrangement known as a governor,[42] in which two masses of mass m are suspended symmetrically from essentially massless rods of length ℓ and rotate around a vertical shaft at angular velocity ω. The rods are hinged where they meet the shaft so that, although the masses have to rotate at the same ω as the shaft, the angle θ between the rods and the vertical can change freely. To further simplify things we will restrict our calculations to just one of the masses (the situation shown on the right side of fig. (7.47)).

    (a) Determine the angle θ at which the mass is rotating, as a function of ω.

    (b) Make physical sense of your result for θ in the limit of large ω.

    (c) You should find that there is a minimal value of ω below which your solution for θ doesn't make sense and isn't valid. To what does this cut-off correspond physically, and what happens physically for values of ω less than this cut-off?

    (d) Determine the angular momentum and kinetic energy of the mass as functions of ω.

    (e) As ω increases, how do the angular momentum, kinetic energy, and potential energy of the mass change?

    (f) If there were variations in the angular velocity ω of the shaft, what effect does the governor have on these variations?

    (g) Why would real governors use two masses instead of just one?

[42] For pedagogical purposes we are here greatly simplifying both the engineering and physics of the governor. For a more — *much* more — involved treatment, see R. Ball & P. Holmes, *Dynamical Systems, Stability, and Chaos*, `https://arxiv.org/abs/nlin.CD/0702044`.

42. A uniform solid cylinder of mass m and radius r is gently placed on a level surface. When initially brought into contact with the surface, the cylinder is rotating at angular velocity ω_0 but not translating; after some ugly skidding, it ends up rolling along the surface without slipping.

    (a) Determine the final angular velocity of the cylinder.

    You will want to think in terms of the changes in the cylinder's linear and angular momentum and how these changes are related to the frictional force between the cylinder and the surface and to each other. See the footnote if you need a further hint.[43]

    (b) Prove that your result is valid no matter how the frictional force between the cylinder and the surface varies as a function of time.

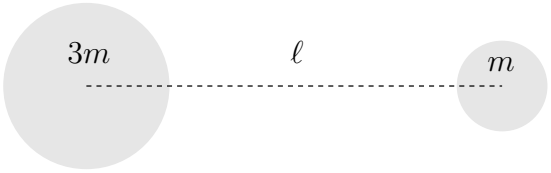


Figure 7.48: Problem 43

43. Although in calculations for gravitational orbits we can often, because the masses are so extremely lopsided, treat the larger mass as "stationary," in a more general and exact treatment the two bodies orbit around their center of mass. Consider two stars, of masses $3m$ and m, in a circular binary orbit around each other, with a distance ℓ between their centers, as shown in fig. (7.48).

    (a) Determine the location of the center of mass of this two-body system.

    (b) Show that $F = ma$ for the circular motion of *both* stars can be written

$$\frac{3Gm^2}{\ell^2} = \frac{3}{4}m\omega^2\ell$$

    (c) This means that ω is the same for both stars. Explain why this had to be so.

    (d) Show that the velocities v of the lighter star and V of the heavier star are then

$$v = \frac{3}{2}\sqrt{\frac{Gm}{\ell}} \qquad\qquad V = \frac{1}{2}\sqrt{\frac{Gm}{\ell}}$$

[43] Think in terms of $\tau = \dfrac{dL}{dt}$ and $F = \dfrac{dp}{dt}$.

(e) Show that the ratio of the angular momenta of the two stars is

$$\frac{L_m}{L_{3m}} = 3$$

(f) Show that the relative velocity of the two stars is

$$v_{rel} = 2\sqrt{\frac{Gm}{\ell}} = \omega\ell$$

(g) By explicitly adding up the individual kinetic energies of the two stars, show that their total kinetic energy is

$$\frac{3}{2}\frac{Gm^2}{\ell}$$

(h) Show that you get this same result for the kinetic energy using eqq. (6.22c) and (6.20) of p.310:

$$K_{rel} = \tfrac{1}{2}\mu v_{rel}^2 \qquad\qquad \mu = \frac{m_1 m_2}{m_1 + m_2}$$

(i) Was it valid to treat the stars as point masses, or are our results above only an approximation that holds when the distance ℓ between the stars is far greater than their radii?

(j) If the radii of the stars are r_{3m} and r_m, should we, by the parallel-axis theorem, use

$$I_{3m} = \tfrac{2}{5}(3m)r_{3m}^2 + (3m)(\tfrac{1}{4}\ell)^2 \qquad\qquad I_m = \tfrac{2}{5}mr_m^2 + m(\tfrac{3}{4}\ell)^2$$

for their moments of inertia?

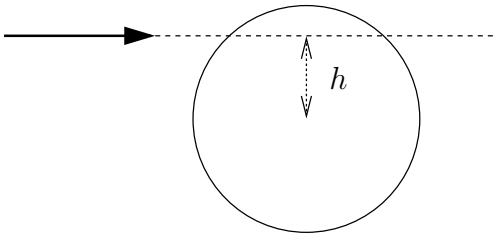


Figure 7.49: Problem 44

44. Determine the height h (see fig. (7.49)) at which you need to strike a cue ball of radius r horizontally with the cue stick so that it never skids — that is, so that it rolls without slipping from the moment that it is struck. Note that although you know nothing about the force that acts during the actual impact, you can set things up in terms of the linear momentum p and angular momentum L imparted to the ball, not unlike what you did in #42. See the footnote if you need a hint.[44]

45. (a) Estimate the magnitude of the Coriolis acceleration for water draining out of a sink. Remember that $\boldsymbol{\omega}$ will correspond to the Earth's rotation about its axis and that $\mathbf{v}$ will correspond to the velocity of the water draining out of the sink. See the footnote if you need a hint.[45]

    (b) You should have found that the acceleration is far too small to account for the swirling of water in drains. What then might account for the swirling?

[44] Start from the condition that corresponds to rolling without slipping and try to relate the quantities in that condition to p and L. The resulting relation between p and L will tell you what you want to know.

[45] You will need to estimate the speed at which the water is heading toward the drain. Also, you can dispense with the cross product by using its maximal possible magnitude.

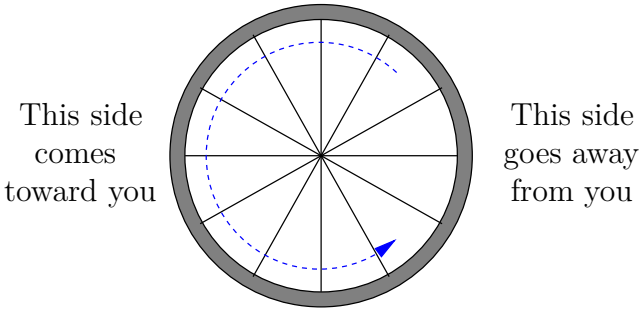


Figure 7.50: Problem 46

46. Fig. (7.50) shows a bicycle wheel rotating counterclockwise in the plane of the page. Imagine that you are holding the wheel by its axle and are going to try to rotate to your left, which will turn the wheel so that the left side will move toward you and the right side away from you, as indicated in fig. (7.50).

    (a) In what direction, as a vector, does the angular momentum of the wheel point?

    (b) In what direction, as a vector, does the *change* that you are trying to impose on the angular momentum of the wheel point?

    (c) What vector torque would be required to produce this change in angular momentum?

    (d) In the absence of such a torque, in what direction will the wheel therefore react (that is, buck)?

47. Except for shotguns, all firearms have rifled barrels: the barrels have grooves that impart a spin to the bullet as it is propelled down the barrel. What is the benefit of putting this spin on the bullet?

7.15 Sketchy Answers

(1(a)i) $\dfrac{8\pi}{T^2}$.

(1(a)ii) $\dfrac{8\pi r}{T}$.

(1(a)iii) $\dfrac{4\pi r}{T}\sqrt{6}$.

(1(a)iv) $T\sqrt{\dfrac{n}{2}}$.

(1(b)i) $\dfrac{v^2}{8\pi r^2}$.

(1(b)ii) $\dfrac{8\pi r}{v}$.

(1(b)iii) $v\sqrt{\frac{3}{2}}$.

(2) $\dfrac{\omega_0^2}{\pi}$.

(3) $\dfrac{2\pi}{\sqrt{Gm_s}}\dfrac{(r_m r_e)^{\frac{3}{2}}}{r_m^{\frac{3}{2}} - r_e^{\frac{3}{2}}} = 199.9\,\text{yr}$.

(4a) $2\pi nf < \omega < (2n+1)\pi f$, $n = 0,\ 1,\ 2,\ \ldots$.

(4b) $(2n+1)\pi f < \omega < 2(n+1)\pi f$, $n = 0,\ 1,\ 2,\ \ldots$.

(6a) The set of answers, not necessarily in order, consists of F_{you} times $\{0,\ \ell,\ \ell\sin\theta,\ 2\ell,\ 2\ell\sin\theta\}$.

(7a) ℓF_{you}.

(7b) $\dfrac{2\ell F_{you}}{mr^2}$.

(7c) $\dfrac{\ell^2 F_{you}^2}{mr^2}\,t^2$.

(7d) $\dfrac{2\ell^2 F_{you}^2}{mr^2}\,t$.

(8) Dude! The answer is already in the problem.

(10(a)i) $\frac{1}{6}\lambda b^2(b+3a)$.

(10(a)ii) $\frac{1}{6}\lambda a^2(a+3b)$.

(10(a)iii) $\frac{1}{6}\lambda(a+b)^3$.

(11a) $33\frac{1}{3}$ rpm. What is the world coming to?

(11c) $\dfrac{m_t}{m_t + m_r}\,\omega_0$.

(12a) $\frac{3}{4}mv^2$ and $\frac{1}{3}$.

(12b) $m'v^2$ and $\frac{1}{2}$.

(15a) $\sqrt{\frac{10}{7}g\ell\sin\theta}$.

(15b) $\frac{5}{7}g\sin\theta$.

(15c) $\frac{2}{7}mg\sin\theta$.

(19a) $\frac{5}{6}g\sin\theta$ and $\dfrac{1}{3}\dfrac{g}{r}\sin\theta$.

(19b) $\sqrt{\frac{5}{3}g\ell\sin\theta}$.

(19c) Distances of $\frac{2}{5}\ell$ and $\frac{3}{5}\ell$ are involved, with $W_f = -\frac{1}{10}mg\ell\sin\theta$.

(20) $a = (1-\kappa)g\sin\theta$ $\qquad$ $\alpha = \dfrac{\kappa}{\beta}\dfrac{g\sin\theta}{r}$

$v = \sqrt{2(1-\kappa)g\ell\sin\theta}$ $\qquad$ $\omega = \dfrac{\kappa}{\beta r}\sqrt{\dfrac{2g\ell\sin\theta}{1-\kappa}}$

distances $\dfrac{\kappa}{\beta(1-\kappa)}\,\ell$ $\qquad$ and $\left(1 - \dfrac{\kappa}{\beta(1-\kappa)}\right)\ell$

$t = \sqrt{\dfrac{2\ell}{(1-\kappa)g\sin\theta}}$ $\qquad$ $W_f = -\kappa mg\ell\sin\theta\left(1 - \dfrac{\kappa}{\beta(1-\kappa)}\right)$

$0 \le \kappa \le \dfrac{\beta}{1+\beta} \le \frac{1}{2}$ $\qquad$ $\frac{1}{2}g\sin\theta \le a \le g\sin\theta$

$0 \le \alpha \le \dfrac{1}{2}\dfrac{g\sin\theta}{r}$

(22a) $\sqrt{\frac{27}{7}gR}$.

(22b) $\sqrt{\frac{27}{7}g(R-r)}$.

(23a) $T_o = \dfrac{m_o(2m_r + \frac{1}{2}m)}{m_o + m_r + \frac{1}{2}m}\,g$ $\quad$ $T_r = \dfrac{m_r(2m_o + \frac{1}{2}m)}{m_o + m_r + \frac{1}{2}m}\,g$

$a = \dfrac{m_o - m_r}{m_o + m_r + \frac{1}{2}m}\,g$ $\quad$ $\alpha = \dfrac{m_o - m_r}{m_o + m_r + \frac{1}{2}m}\,\dfrac{g}{R}$

(23d) $\sqrt{\dfrac{2gh\,|m_r - m_o|}{m_o + m_r + \frac{1}{2}m}}$.

(24a) $a = \dfrac{1}{1 + I/mr^2}\,g$, $\quad T = mg\,\dfrac{I/mr^2}{1 + I/mr^2}$.

(25a) $\dfrac{m_c + m_p}{m_c + \frac{3}{2}m_p}\,g$.

(26(b)i) $a = \dfrac{2T}{3m}, \quad F_f = \frac{1}{3}T.$

(26(b)ii) $a = \dfrac{4T}{3m}, \quad F_f = \frac{1}{3}T.$

(27a) $\dfrac{m_f}{m_f + \frac{1}{2}(m_c + m_p)}\, g.$

(27c) The cat's and cylinder's accelerations, cylinder's angular acceleration, force of friction between the cylinder and the surface, and tensions in the upper and lower parts of the cord are

$$a_f = \frac{m_f g}{m_f + \frac{1}{2}m_p + \frac{3}{8}m_c} \qquad a_c = \frac{\frac{1}{2}m_f g}{\left(m_f + \frac{1}{2}m_p + \frac{3}{8}m_c\right)}$$

$$\alpha_c = \frac{\frac{1}{2}m_f g}{\left(m_f + \frac{1}{2}m_p + \frac{3}{8}m_c\right) r_c} \qquad F_f = \frac{\frac{1}{8}m_f m_c g}{m_f + \frac{1}{2}m_p + \frac{3}{8}m_c}$$

$$T_u = \frac{\frac{3}{8}m_f m_c g}{m_f + \frac{1}{2}m_p + \frac{3}{8}m_c} \qquad T_\ell = \frac{m_f g\left(\frac{1}{2}m_p + \frac{3}{8}m_c\right)}{m_f + \frac{1}{2}m_p + \frac{3}{8}m_c}$$

(28a) $T = \dfrac{mM}{m+3M}\, g\sin\theta, \quad a = \dfrac{2M}{m+3M}\, g\sin\theta, \quad F_f = \dfrac{M^2}{m+3M}\, g\sin\theta.$

(29a) $\sqrt{2g\ell\cos\theta}.$

(29b) $3mg.$

(29c) $a_{tan} = -g\sin\theta, \quad a_r = -2g\cos\theta.$

(30a) $\sqrt{3g\ell\cos\theta}.$

(30b) $\sqrt{2g\ell\cos\theta\left(\dfrac{m+\frac{1}{2}m'}{m+\frac{1}{3}m'}\right)}.$

(31a) $\sqrt{3g\ell}.$

(31c) $a_r = -3g(1-\cos\theta), \quad a_{tan} = \frac{3}{2}g\sin\theta.$

(31d) $-\frac{3}{2}(1 + 2\cos\theta - 3\cos^2\theta)g.$

(32a) $\sqrt{2g\ell\cos\theta_0}.$

(32b) $\sqrt{\dfrac{2g\ell}{\cos\theta_0}}.$

(34a) $K = m\omega_0^2\ell^2$ and $L = 2m\omega_0\ell^2.$

(34b) $\omega = \dfrac{\omega_0}{\kappa^2}$, $K = \dfrac{m\omega_0^2\ell^2}{\kappa^2}$, and $L = 2m\omega_0\ell^2.$

(34c) $\dfrac{m\omega_0^2\ell}{\kappa^3}$.

(35a) The point is $\frac{1}{4}\kappa\ell$ above the dotted line in fig. (7.41).

(35b) $\frac{1}{12}m\ell^2(1+\frac{3}{2}\kappa^2)$.

(35c) $\frac{1}{2}v_0$.

(35e) $\dfrac{1}{1+\frac{3}{2}\kappa^2}\left(\omega_0 - 3\kappa\dfrac{v_0}{\ell}\right)$.

(36(b)ii) Depending on your sign convention, $\pm\dfrac{3v_0}{4\ell} \mp \dfrac{\omega_0}{8}$.

(36(b)iii) $-\frac{1}{192}m\ell^2\left(7\omega_0^2 + 12\omega_0\dfrac{v_0}{\ell} + 12\dfrac{v_0^2}{\ell^2}\right)$.

(38b) $\frac{1}{2}v_0$.

(38d) mv_0R.

(38e) $\frac{1}{2}mv_0R$.

(38f) $\frac{1}{2}mv_0R$.

(39) $\omega = \omega_0$ and $v = \frac{1}{4}\omega_0\ell$.

(40a) $\dfrac{3v_0}{4\ell}$.

(40b) $\frac{1}{8}mv_0^2$.

(40c) $\frac{1}{2}v_0$.

(40d) $\frac{9}{16}v_0$.

(41a) $\cos\theta = \dfrac{g}{\omega^2\ell}$.

(41d) $L = m\omega\ell^2\left(1 - \left(\dfrac{g}{\omega^2\ell}\right)^2\right)$, $\quad K = \frac{1}{2}m\omega^2\ell^2\left(1 - \left(\dfrac{g}{\omega^2\ell}\right)^2\right)$.

(42a) $\frac{1}{3}\omega_0$.

(44) $\frac{2}{5}r$.

Chapter 8

Static Equilibria

8.1 The Conditions of Equilibrium

$F = ma$ tells us that a body will move if there is a net force on it. Similarly, $\tau = I\alpha$ tells us that a body will rotate if there is a net torque on it. If a body *remains* motionless, the net force and the net torque exerted on it must therefore both vanish, or else the body would begin to translate or rotate. Such a body is said to be in static equilibrium.

Note that it is critically important that the body be in this motionless state for an extended time; it is not enough that the body be motionless for an instant. When you throw a ball straight up, for example, the ball experiences the downward acceleration g due to gravity throughout its entire flight: on the way up, on the way down, and also at the moment that it is at the apex — where it is, for an instant, at rest. The term static generally means not just at rest, but *remaining at rest*.

If we know that a body is in static equilibrium, we therefore also know that

$$\mathbf{F}_{net} = 0 \quad \text{and} \quad \boldsymbol{\tau}_{net} = 0$$

for the body. This gives us a means of determining the forces supporting the body: for the kinds of effectively two-dimensional cases that will concern us, we draw the force diagram for the body, then balance the forces and torques on the body by setting the net vertical force, the net horizontal force, and the net torque all to zero.

In two dimensions, the torques on the body will all be of the usual form $\tau = r_\perp F$, where $r_\perp$ is the torque arm. Usually we measure the torque arm from the axis of rotation to the line of force, but how do you do that when the body *isn't* rotating and there consequently isn't any axis of rotation? It turns out, as we will prove below, that when there is no net force on the body, you can evaluate your torques about any axis you want, as long as you are consistent and always use that same axis. In practice, you try to be as

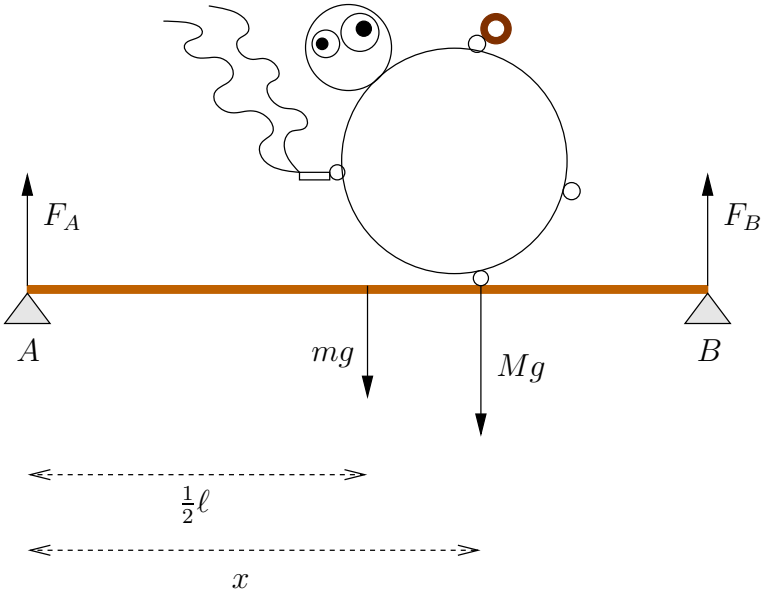


Figure 8.1: What's Wrong with This Picture?

lazy as possible[1] and choose the axis that will make the torques easiest to evaluate. Often this will mean choosing a point where one or more forces act, so that the torque arms for those forces and hence the torques due to them vanish.

Once you've chosen the axis for evaluating the torques, remember also that you must be careful to keep your $\pm$ signs straight for clockwise and counterclockwise. You should also be mindful that, as was shown in §7.9, the force mg of gravity effectively acts at the center of mass of the body.

An Example

To pick a mundane but classic example, suppose we have a chain-smoking ballerina/o with a passion for little chocolate donuts working out on a balance beam. The beam, of mass m and length ℓ, is supported at its two extreme ends, A and B, as shown in fig.(8.1). If the beam is uniform, its center of mass, and hence the place where the weight mg of the beam effectively acts, will be at its geometric center. The ballerina/o is standing on one foot a distance x from end A and exerts on the beam a downward normal force that will be equal to his or her weight Mg. The two supports exert unknown upward forces F_A and F_B on the beam.

In this case, all the forces are vertical, so we get only one relation from

[1] We expect that this will have wide appeal. ☺

$\mathbf{F} = m\mathbf{a}$ — the one for the components along the vertical direction:

$$F_A + F_B - mg - Mg = ma = 0 \tag{8.1}$$

Note that, since we are setting up $F = ma$ for the beam, we have used the beam's mass m on the ma side. Since the beam is static, $a = 0$. Eq. (8.1) is only one equation in the two unknowns F_A and F_B, so we can't solve it, but we can rewrite it as

$$F_A + F_B = mg + Mg = (m + M)g$$

Although we cannot yet solve for the forces F_A and F_B separately, this tells us that together they support the combined weight of the ballerina/o and the beam. The question is how that total weight is divided between them.

If we take our torques about end A, then F_A has zero torque arm and hence zero torque. For mg, the torque arm is $\frac{1}{2}\ell$, and if we take clockwise to be positive, the corresponding torque is thus

$$\tfrac{1}{2}\ell mg$$

For Mg, the torque arm is x, and the corresponding clockwise torque is thus

$$xMg$$

Finally, for F_B, the torque arm is ℓ, and the corresponding counterclockwise torque is

$$-\ell F_B$$

Balancing the torques on the beam therefore gives us

$$\begin{aligned} \tau &= I\alpha \\ \tfrac{1}{2}mg\ell + Mgx - F_B\ell &= 0 \\ F_B &= \tfrac{1}{2}mg + Mg\frac{x}{\ell} \end{aligned}$$

Using this result for F_B in eq. (8.1) then gives

$$F_A = \tfrac{1}{2}mg + Mg\frac{\ell - x}{\ell}$$

Thus supports A and B divide the weight of the beam, which effectively acts halfway between them at the center of the beam, equally. But how they share the weight of the ballerina/o depends on where the ballerina/o is standing: the more biased the ballerina/o's location, the more unequally his or her weight is divided between the supports. A supports a fraction $(\ell - x)/\ell$ of the ballerina/o's weight, B the remaining fraction x/ℓ. In the

extreme case $x = 0$, when the ballerina/o is standing directly over support A, all of his or her weight is supported by A. Likewise in the extreme case $x = \ell$, when the ballerina/o is standing directly over support B, all of his or her weight is supported by B. And when the ballerina/o is smack in the middle of the beam ($x = \frac{1}{2}\ell$), his or her weight is equally divided between A and B.

Now that we've seen how neatly that all worked out, we might go back and question our assumption that the forces exerted on the beam by the supports A and B were vertical: Why can't F_A and F_B also have horizontal components? The answer, of course, is that they can: depending on the construction of the beam and the way the beam was placed or mounted on the floor, there may very well be a tension or compression exerted on the beam by its supports. This would, however, introduce two additional unknowns (the horizontal components of F_A and F_B) but only one additional equation (the horizontal component of $\mathbf{F} = m\mathbf{a} = 0$) and would thus leave us with an indeterminate system: we would know that the horizontal forces exerted on the beam by the supports would have to be equal and opposite, but without more information we could not determine their magnitude or even whether they were a tension or a compression. The moral of all this is that real-world engineering problems are often very messy and difficult. For the most part we will, however, follow the custom of introductory textbooks and steer well clear of reality, restricting ourselves to more or less bogus situations where the number of equations and the number of unknowns are conveniently if rather cynically commensurate.

A Proof

We now prove that when there is no net force on the body, you will get the same result for the net torque about any axis.

Suppose that a set of forces $\{\mathbf{F}_i\}$, $i = 1, 2, \ldots, n$, act on a body. If we evaluate all our torques about axis A in fig. (8.2), then the torque due to force $\mathbf{F}_i$ is

$$\boldsymbol{\tau}_i = \mathbf{r}_i \times \mathbf{F}_i$$

and the net torque on the body is just the sum of these torques:

$$\boldsymbol{\tau}_{net} = \sum_{i=1}^{n} \boldsymbol{\tau}_i = \sum_{i=1}^{n} \mathbf{r}_i \times \mathbf{F}_i \tag{8.2}$$

If instead we evaluate all our torques about axis A' in fig. (8.2), then the torque due to force $\mathbf{F}_i$ is

$$\boldsymbol{\tau}_i{}' = \mathbf{r}_i' \times \mathbf{F}_i$$

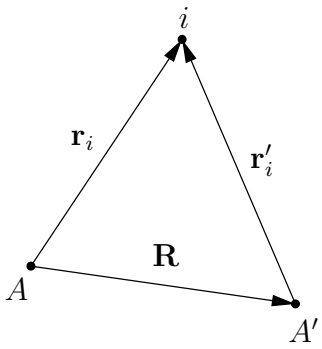


Figure 8.2: Torques About Two Different Axes

and the net torque on the body is

$$\boldsymbol{\tau}'_{net} = \sum_{i=1}^{n} \boldsymbol{\tau}_i{}' = \sum_{i=1}^{n} \mathbf{r}'_i \times \mathbf{F}_i \tag{8.3}$$

As you can see from fig. (8.2),

$$\mathbf{r}_i = \mathbf{R} + \mathbf{r}'_i$$

where $\mathbf{R}$ is the displacement vector that goes between the axes A and A'. Using this in eq. (8.2), we have

$$\begin{aligned} \boldsymbol{\tau}_{net} &= \sum_{i=1}^{n} (\mathbf{R} + \mathbf{r}'_i) \times \mathbf{F}_i \\ &= \sum_{i=1}^{n} \mathbf{R} \times \mathbf{F}_i + \sum_{i=1}^{n} \mathbf{r}'_i \times \mathbf{F}_i \end{aligned}$$

The second sum is just the result we had in eq. (8.3) for the net torque $\boldsymbol{\tau}'_{net}$ about axis A'. In the first sum, the $\mathbf{R}$ is constant and can be pulled outside the sum. Thus

$$\boldsymbol{\tau}_{net} = \mathbf{R} \times \sum_{i=1}^{n} \mathbf{F}_i + \boldsymbol{\tau}'_{net}$$

The sum over the $\mathbf{F}_i$ is just the net force on the system. If there is no net force on the system, then

$$\boldsymbol{\tau}_{net} = \boldsymbol{\tau}'_{net}$$

That is, if there is no net force on the system, we get the same result for the net torque about any two axes. Word.

8.2 Stable & Unstable Equilibria

Static equilibria are of two types: stable and unstable. When a body is in a stable equilibrium, it will return to that equilibrium when it experiences

small perturbations, that is, small nudges away from equilibrium. A marble in the bottom of a bowl is in a stable equilibrium: if you give the marble a little tap, it will roll slightly up the side of the bowl, come to rest, then roll back down toward the bottom of the bowl.[2] When a body is in an unstable equilibrium, a small perturbation will result in a runaway effect that takes it farther and farther away from the equilibrium point. A marble balanced on top of an inverted bowl is in an unstable equilibrium: a little tap, no matter how slight, will result in the marble rolling off the bowl.

Since a condition of equilibrium is that the net force on the body vanish, and since force is related to potential energy by eq. (5.13) of p.264,

$$F = -\frac{dU}{dx}$$

or, in higher dimensions, eq. (5.14),

$$\mathbf{F} = -\nabla U$$

we can look at equilibria in terms of potential energy as well as in terms of forces. In one dimension, $F = 0$ means $dU/dx = 0$, that is, that we are at a flat spot on the potential-energy curve. For a stable equilibrium, we want the force F and the perturbation dx to be in opposite directions, so that whichever way you perturb the body, the force is back toward the equilibrium point. In one dimension, having opposite directions means having opposite signs, so that we want the product $F\,dx$ to be negative. If we rewrite eq. (5.13) in the form

$$F\,dx = -dU$$

we see that negative $F\,dx$ corresponds to positive dU: for a stable equilibrium, dU must be positive when we move away from the equilibrium point, so that we must be in a "valley" of the potential-energy curve. Mathematically, being in a valley corresponds to being at a minimum, and being at a minimum of the function U means

$$\frac{dU}{dx} = 0 \qquad \text{with} \qquad \frac{d^2U}{dx^2} > 0$$

For unstable equilibria it is just the opposite: a runaway force means that F and dx are in the same direction, so that $F\,dx$ is positive and dU negative, which in turn corresponds to being at a "peak" of the potential-energy curve. Mathematically being at a peak corresponds to being at a maximum, and being at a maximum of the function U means

$$\frac{dU}{dx} = 0 \qquad \text{with} \qquad \frac{d^2U}{dx^2} < 0$$

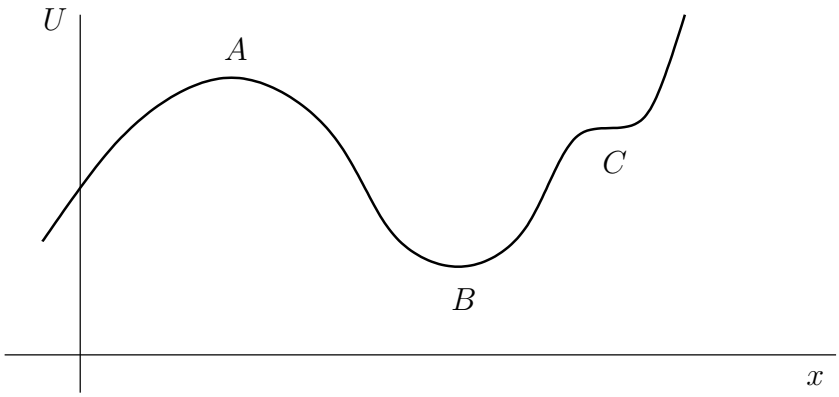


Figure 8.3: Stable & Unstable Equilibria

In fig. (8.3), point A is an unstable equilibrium point, point B a stable equilibrium point, and point C a pathological case ($dU/dx = 0$ with $d^2U/dx^2 = 0$) that would also have to be considered unstable: if the marble were nudged to the right, it would return to the equilibrium point, but would then run away to the left.

In higher dimensions, it is also possible to have saddle points — equilibrium points that are stable along some axes but unstable along others. If you think about the shape of an actual saddle, this name makes perfect sense: along the front-back axis, the saddle is bowl-shaped, so that a marble would be in a stable equilibrium along this axis. But along the side-to-side axis, the saddle is shaped like an inverted bowl, so that the marble would be in an unstable equilibrium along that axis.

[2]Of course, if you really whack it, the marble will fly up out of the bowl, but such a perturbation, while much more entertaining, no longer counts as small.

8.3 Problems

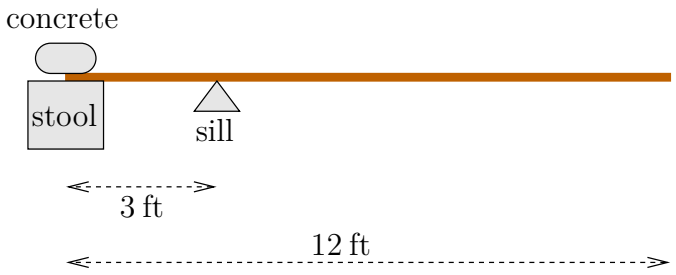


Figure 8.4: Problem 1

1. You put on an eye patch, tie a bandanna around your head, lay a 12 ft piece of (by comparison, massless) pine shelving across a second-floor window sill, and weigh the inner end down on top of a stool with an 80 lb bag of concrete, as sketched in fig. (8.4). You then take a pitchfork and force your 60 lb little brother/sister to "walk the plank." Bwahahaha!

   (a) How far out from the sill does the sibling make it?

   (b) How would you need to shift the concrete and plank so that, for theatrical effect, the plank gives way just as the sibling makes it to the extreme end?

2. Our chain-smoking ballerina/o with a passion for little chocolate donuts (mass m — the ballerina/o, that is, not the donuts) has just had a smoke and a donut in the alleyway and is back on the beam.[3] The beam has length ℓ and mass m', but this time its two supports are "indented" a distance a from each end.

   (a) Determine the forces exerted by the supports as a function of the distance x that the ballerina/o is standing from the center of the beam.

   (b) Make physical sense of your result for these forces when

      i. $x = 0$.

      ii. $x = \frac{1}{2}\ell - a$.

      iii. $x = \frac{1}{2}\ell$.

   (c) Could one or the other of the supporting forces be negative in any of the above cases? If so, how would you interpret this physically? How would the beam have to be constructed to allow for such a force?

[3] Okay, so the repetition is a little lame. But how much can you do with a balance beam? We have to have *something* to work with.

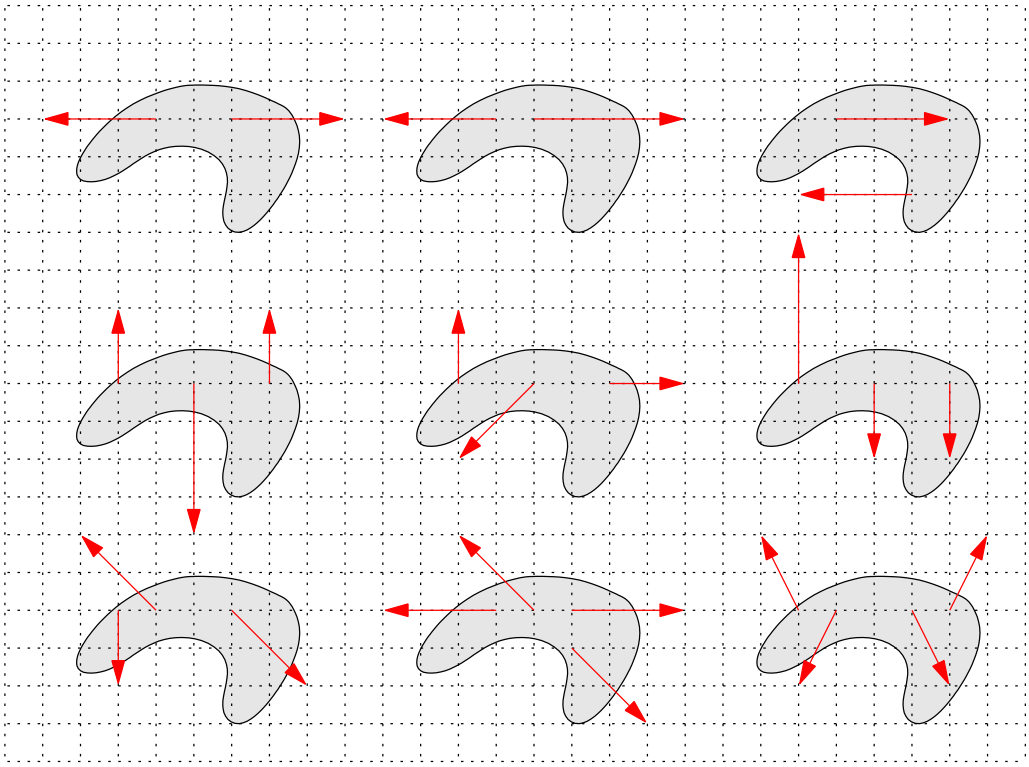


Figure 8.5: Problem 3

3. Fig. (8.5) shows the rather funky body — over and over again, sort of like the spatial equivalent of *Groundhog Day*. In different cases, different forces act on the body, but all the forces are drawn both at the points where they are exerted on the body and to scale against the background grid.

   (a) Determine for each case

      i. Whether the forces on the body are balanced.

      ii. Whether the torques on the body are balanced.

   (b) If for a particular case the body is not in equilibrium,

      i. What minimal number of additional forces is needed to produce equilibrium?

      ii. Determine at least one such set of forces.

   (c) Are any of your answers to $^{\#}3$(b)ii unique?

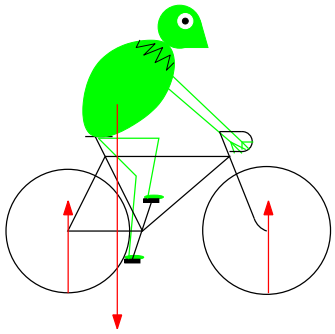


Figure 8.6: Problem 4

4. Fig. (8.6), as you can plainly see, shows Kermit the Frog on a bicycle. The upward red arrows represent the supporting forces exerted on the bicycle by the ground, and the downward red arrow, drawn from the center of mass of the rider-bicycle system, represents the weight of the system. Note that this latter force acts closer to the rear than to the front wheel. Although bicycles are not normally stationary, when you are coasting your acceleration and angular acceleration are both zero, so that forces and torques are balanced just as they would be in static equilibrium. Such situations, which are equivalent to static equilibria even though there is motion, are known as quasistatic.

   (a) What can you conclude about the relative magnitudes of the supporting forces acting at the front and rear wheels? Would you therefore expect that flats are more common in front or rear tires?

   (b) Suppose now that you are pedaling hard, so that at the point of contact between the rear wheel and the ground there is a substantial horizontal frictional force exerted on the bike in the forward direction. How does this change the relative magnitudes of the vertical forces acting at the front and rear wheels?

   (c) Suppose now that you are applying the rear brake hard, so that at the point of contact between the rear wheel and the ground there is a substantial horizontal frictional force exerted on the bike in the backward direction. How does this change the relative magnitudes of the vertical forces acting at the front and rear wheels? To what familiar consequences will this lead?

   (d) Suppose now (sure seems to be an awful lot of supposing going on around here) that you are applying the front brake hard, so that at the point of contact between the front wheel and the ground there is a substantial horizontal frictional force exerted on the bike in the backward direction. How does this change the relative magnitudes of the vertical forces acting at the front and rear wheels? To what familiar consequences will this lead?

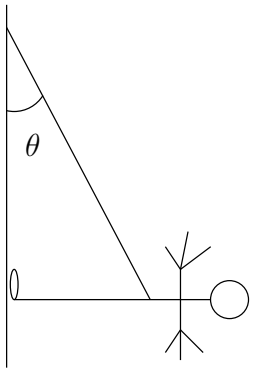


Figure 8.7: Problem 5

5. Now that we know about torques as well as forces, we will revisit #27 of p.234, which was about rappelling. In our last episode, you found yourself (mass m) in the ridiculous situation shown in fig. (8.7), with your legs and body at a right angle to the rock face. Let us now suppose that the length of rope running along the hypotenuse of the triangle is ℓ and that your end of the rope is "connected" to your center of mass.

   (a) If there is friction between your feet and the rock face,

      i. What can you conclude about the horizontal and vertical forces the rock face exerts on you?

      ii. Is it possible to solve for the forces acting on you? If so, solve for them. If not, explain why not.

   (b) If the rock face is instead frictionless,

      i. What can you conclude about the horizontal and vertical forces the rock face exerts on you?

      ii. Is it possible to solve for the forces acting on you? If so, solve for them. If not, explain why not.

6. Explain in terms of torques how crowbars and other levers work.

7. Certain kinds of witch doctors sometimes check for an imbalance in posture by having you stand on two scales, with one foot on each scale. Is there any validity to such a test, and what would it show?

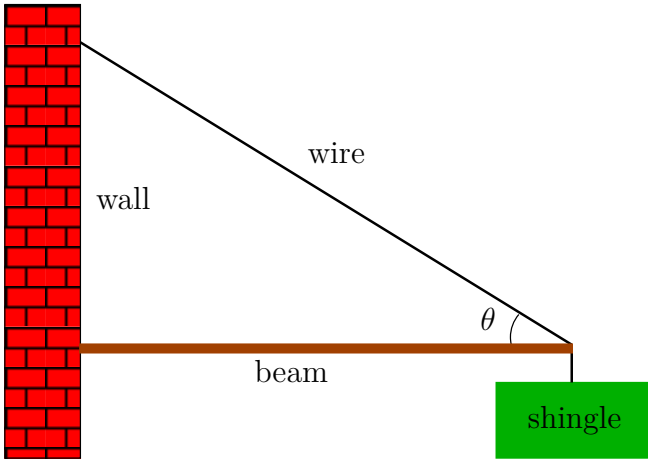


Figure 8.8: Problem 8

8. And now for the classic shingle[4] problem: in fig. (8.8), the uniform beam has mass m_b and length ℓ, the shingle has mass m_s, and the supporting wire is at angle θ to the horizontal.

   (a) Determine all of the forces acting on the beam. See the footnote if you need a hint.[5]

   (b) Make physical sense of the behavior of your results as functions of θ, including the extreme limits

      i. $\theta \to \frac{\pi}{2}$.

      ii. $\theta \to 0$.

[4] *Shingle*, for those of you unfamiliar with this usage, is a nearly obsolete word for *sign*; nowadays, it is used mostly in two other senses having to do with roofing and a medical condition that makes old people grumpy.

[5] At the juncture of the beam with the wall, you must allow that the wall may exert both a horizontal and a vertical force on the beam.

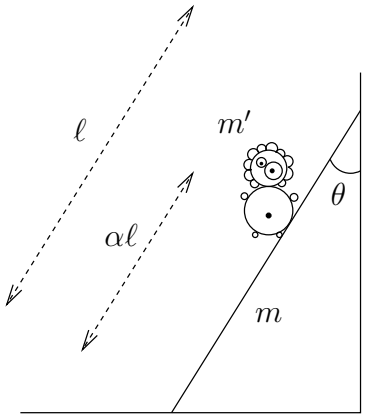


Figure 8.9: Problem 9

9. Fig. (8.9) shows a uniform ladder of length ℓ and mass m leaned at angle θ against a wall. There is friction between the ladder and the ground, but not between the ladder and the wall.[6] An unattended baby of mass m' has climbed a distance $\alpha\ell$ up the ladder, as measured along the ladder from its base, where $0 \leq \alpha \leq 1$. (Although it is obviously not a very good approximation, we will take the baby to act on the ladder in a point-like way at exactly $\alpha\ell$ from its base.)

   (a) Determine all the forces acting on the ladder. See the footnote if you need a hint.[7]

   (b) Make physical sense of the behavior of these forces as functions of α and θ. Be sure to consider also the extreme values of α and θ.

[6] With real ladders, there is of course substantial friction between the ladder and the wall, but allowing for this additional force on the ladder would render the problem insoluble without further information. This is one of those cases where reality is inconveniently complicated.

[7] While the force exerted by the ground on the base of the ladder may have both a vertical and a horizontal component, the absence of friction with the wall means that the force that the wall exerts on the ladder can have only a component perpendicular to the wall.

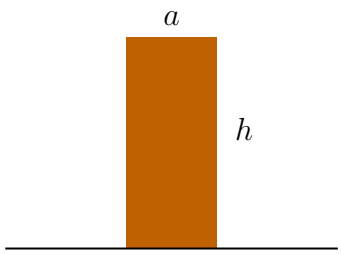


Figure 8.10: Problem 10

10. Fig. (8.10) shows an end view of a bookcase of depth a and height h. There is nothing special about this bookcase. And all it's doing is just sitting there on the level floor of a quiet suburban library. Sorry to have gotten your hopes up.

    Anyway, suddenly a bunch of Hell's Angels intent on mayhem bursts into the library and starts doing physics experiments on the furniture.

    (a) If we treat the mass m of the bookcase as though it were evenly distributed throughout its volume, how far can the bookcase be tipped before it will fall over? Express your answer as an angle. See the footnote if you need a hint.[8]

    (b) Will there be a need for a horizontal frictional force (or other horizontal force) to prevent the bookcase from slipping along the floor as it is falling?

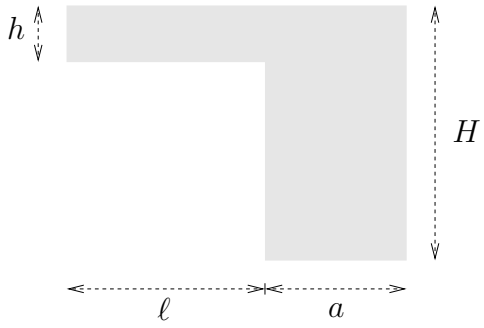


Figure 8.11: Problem 11

11. Commissioned to create a monument to stupidity, you devise the Stone Henge-like design shown, in cross section, in fig. (8.11). The monument is to be carved out of a single piece of solid granite. Taking the values of a, ℓ, and H to be fixed, what is the largest value h can have without the monument falling over and squashing those admiring it? (Not that that would necessarily be a bad thing.) See the footnote if you need a hint.[9]

[8] Think about where the center of mass of the bookcase is located relative to the pivot point.

[9] Again, think about where the center of mass is located relative to the pivot point.

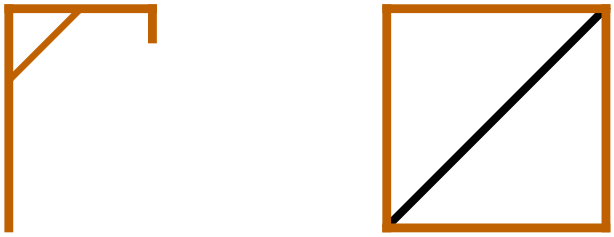


Figure 8.12: Problem 12

12. (a) The left side of fig. (8.12) shows the outline of a classic gallows. What is the purpose of the diagonal piece in the gallows? In terms of forces and torques, how does this piece make the gallows stronger?

    (b) The right side of fig. (8.12) shows the outline of a classic fence gate. Sometimes the diagonal piece is wooden, sometimes it is a cable with a turnbuckle to adjust its tension. Which side of the gate should be the hinge side if the diagonal piece is wooden? If the diagonal piece is a cable?

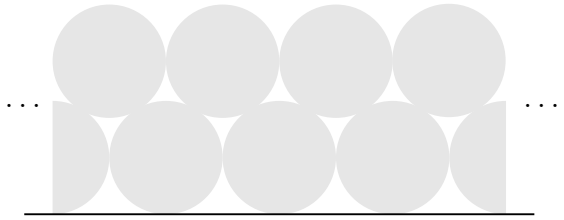


Figure 8.13: Problem 13

13. Fig. (8.13) shows an end view of a two-layer stack of pipes, each of mass m and radius r, on a level surface. The dots are supposed to indicate that this arrangement continues, as far as we are concerned, out to infinity on both the right and the left sides. We will assume that there is no friction between pipes or between the pipes and the surface.

    (a) Determine as many of the forces acting on each pipe as you can.

    (b) Were there any forces for that you were unable to determine? If so, what would determine these forces? That is, what conditions or additional information would allow us to determine the values of these forces?

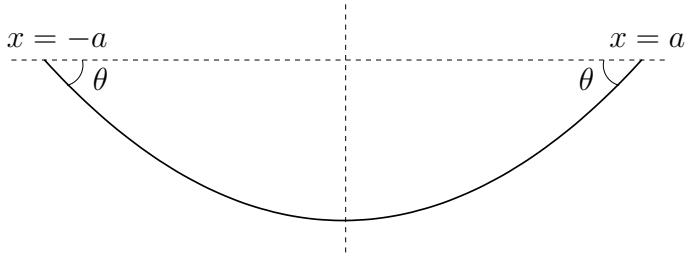


Figure 8.14: Problem 14

14. Recall the catenary from §4.8: in the symmetric case shown in fig. (8.14), the ends of a hanging cord of mass m are located at $x = -a$ and $x = a$. At each end, the tangent to the rope makes an angle θ with the horizontal.

    (a) Determine the tension at the ends of the cord as a function of θ.

    (b) Show that this result is consistent with the results of §4.8 (eqq. (4.20) and (4.21)) for the geometric shape $y(x)$ and length ℓ of cord in the catenary:

$$y = \frac{T_x}{\lambda g}\cosh\left(\frac{\lambda g}{T_x}\,x\right) \qquad\qquad \ell = 2\,\frac{T_x}{\lambda g}\,\sinh\left(\frac{\lambda g}{T_x}\,a\right)$$

    where λ is the linear density (mass per unit length) of the cord and T_x is the (constant) horizontal tension. It will help to recall that the angle θ between the tangent to a curve $y(x)$ and the horizontal is given by $\tan\theta = dy/dx$, that

$$\sinh\alpha x = \frac{e^{\alpha x} - e^{-\alpha x}}{2} \qquad\qquad \cosh\alpha x = \frac{e^{\alpha x} + e^{-\alpha x}}{2}$$

    and that therefore

$$\frac{d}{dx}(\cosh\alpha x) = \alpha\sinh\alpha x$$

8.4 Sketchy Answers

(1a) 4 ft.

(1b) Pull the plank in until the sill is $\frac{36}{7}$ ft from concrete.

(2a) $\frac{1}{2}(m+m')g \pm \dfrac{\frac{1}{2}x}{\frac{1}{2}\ell - a}\, mg.$

(8a) The set of forces is

$$\{m_b g,\ m_s g,\ \tfrac{1}{2}m_b g,\ (m_s + \tfrac{1}{2}m_b)\, g \csc\theta,\ (m_s + \tfrac{1}{2}m_b)\, g \cot\theta\}$$

(9a) The set of forces is $\{mg,\ m'g,\ (m+m')g,\ (\alpha m' + \frac{1}{2}m)g\tan\theta\}$.

(10a) The angle with the vertical is given by $\tan\theta = \dfrac{a}{h}$.

(11) $\left(\dfrac{a}{\ell}\right)^2 H$.

(13a) The set of determinable forces is $\{mg,\ \frac{1}{\sqrt{3}}mg,\ 2mg\}$.

(14a) $\dfrac{mg}{2\sin\theta}$.

Chapter 9

Harmonic Motion

9.1 The First Refrain

Mathematically, harmonic motion is defined as motion that is sinusoidal, that is, motion for which the plot would have the shape of a sine wave. Physically, the simplest example of harmonic motion is the oscillation of a mass on the end of a spring. Only a very few parameters (all of which are already familiar to you) are needed to describe such motion:

- The period T is the time for each full cycle, that is, the time to go from one endpoint to the other *and back again.*
- The frequency f is the reciprocal of the period, $f = 1/T$: it is a rate that indicates, in cycles per unit time, how rapidly the mass is oscillating.
- The angular frequency ω is given by $\omega = 2\pi f$ and also indicates the rate at which the mass is oscillating, but in terms of angle (radians per unit time) rather than cycles per unit time.
- The amplitude A is the distance from the midpoint to either endpoint, as shown in fig. (9.1).

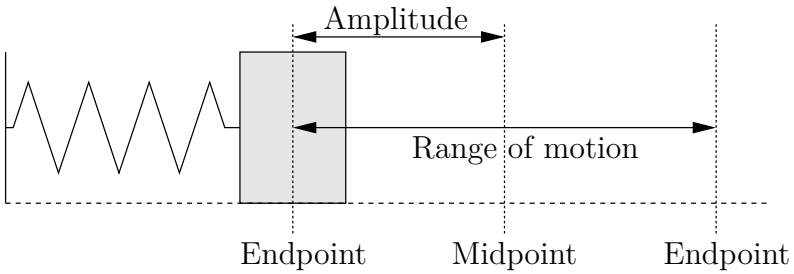


Figure 9.1: Oscillation of a Mass on a Spring

Suppose we set up our x axis along the oscillation with the origin $x = 0$ at its midpoint. The force that the spring exerts on the mass is then described by the usual spring force relation $F = -kx$. Since this spring force is symmetric about the midpoint $x = 0$, so is the resulting motion. If we use $F = -kx$ in $F = ma$ and note that $a = d^2x/dt^2$, we have

$$-kx = m\frac{d^2x}{dt^2} \tag{9.1}$$

Eq. (9.1) is a very common and relatively simple differential equation for x as a function of the time t. The general solution to this differential equation is the harmonic oscillation

$$x = A\cos(\omega t + \varphi) \tag{9.2}$$

where A is the amplitude, ω is the angular frequency, and the new animal φ, which corresponds to the alignment (the "horizontal shift") of this cosine wave along the t axis, is called the phase. If you've had enough math, you already know that eq. (9.2) is the solution to eq. (9.1). If you haven't, we're about to verify that this solution works, anyway: if we plug this solution for x into eq. (9.1), we have

$$\begin{aligned} -kx &= m\frac{d^2}{dt^2}\Big(A\cos(\omega t + \varphi)\Big) \\ &= -m\omega^2 A\cos(\omega t + \varphi) \\ &= -m\omega^2 x \end{aligned}$$

which is true as long as we take ω to be[1]

$$\omega = \sqrt{\frac{k}{m}} \tag{9.3}$$

[1]While we have now shown that solution (9.2) satisfies eq. (9.1), we have not shown that (9.2) is the *only* solution to eq. (9.1). This turns out, however, to be the case: as we will see in §20.1, linear differential equations are equivalent to algebraic equations of the same order, so that our second-order differential eq. (9.1) is equivalent to a quadratic equation. This means that, corresponding to there being two and only two roots to a quadratic equation, eq. (9.1) will have two and only two functionally independent solutions. Our solution (9.2) looks like just a single function, but because of the parameter φ the $A\cos(\omega t + \varphi)$ is actually a sine and a cosine rolled together:

$$\begin{aligned} x &= A\cos(\omega t + \varphi) \\ &= A(\cos\omega t\cos\varphi - \sin\omega t\sin\varphi) \\ &= A\cos\varphi\cos\omega t - A\sin\varphi\sin\omega t \\ &= A'\cos\omega t + B'\sin\omega t \end{aligned}$$

where $A' = A\cos\varphi$ and $B' = -A\sin\varphi$ are two new parameters equivalent to A and φ.

Some things to note about the $x(t)$ given by eq. (9.2):

- The cosine function is indeed sinusoidal (harmonic), as advertised.
- As the cosine varies from $+1$ to -1, x varies from $+A$ to $-A$. That is, the oscillation extends out to one amplitude on either side of the midpoint, as advertised.
- The angle inside the cosine is $\omega t + \varphi$, which means that ω is indeed the angular rate (that is, the angular frequency) of the oscillation, as advertised.
- The phase φ, as mentioned above, corresponds to the alignment (the "horizontal shift") of this cosine wave along the t axis: physically, the mass need not be $x = +A$ initially, and mathematically it is the phase φ that allows it to be anywhere in the cycle at time $t = 0$.

  It turns out, however, that in many applications we are not all that concerned with the phase, so for simplicity we will often assume that $\varphi = 0$. When we set $\varphi = 0$, our solution for the oscillation of a mass on a spring reduces to

$$x = A\cos\omega t \qquad \text{with} \qquad \omega = \sqrt{\frac{k}{m}} \tag{9.4a}$$

Taking time derivatives of eq. (9.4a), we obtain corresponding results for the velocity and acceleration:

$$v = \frac{dx}{dt} = \frac{d}{dt}(A\cos\omega t) = -\omega A\sin\omega t \tag{9.4b}$$

$$a = \frac{dv}{dt} = \frac{d}{dt}(-\omega A\sin\omega t) = -\omega^2 A\cos\omega t \tag{9.4c}$$

As functions of time, x, v, and a curves are therefore like a cosine, an upside-down sine, and an upside-down cosine, respectively, as shown for one full cycle in fig. (9.2), in which the red line corresponds to x, the green line to v, and the blue line to a. Now, x, v, and a all have different physical dimensions, of course, so that the dimensions of the vertical axes of their plots as functions of time differ and they cannot sensibly be plotted together on a single pair of axes. In the superposition of their plots in fig. (9.2), no meaningful comparison of the "vertical" values can therefore be made among x, v, and a. Since x, v, and a are, however, all parametrized by the same time t, fig. (9.2) is nonetheless useful for seeing the *temporal* relationships among them, and there are several observations to be made about these relationships:

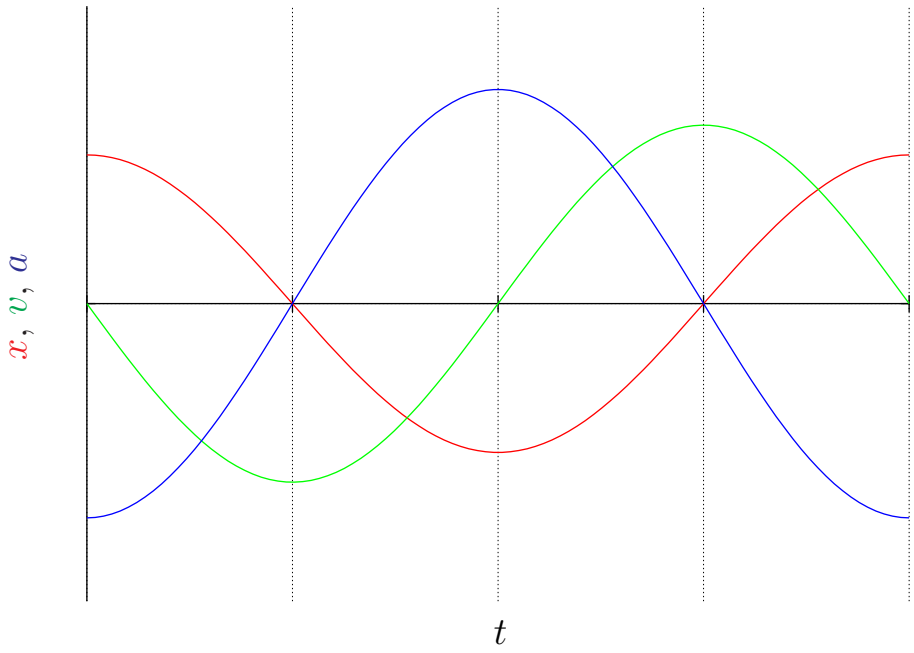


Figure 9.2: Almost Complete Nonsense

- From eqq. (9.4a) and (9.4c) we see that $a = -\omega^2 x$, that is, that a is proportional to x, with a negative constant of proportionality $-\omega^2$. The plots of x and a thus mirror each other: when one is at its maximum, the other is at its minimum, and the curves are always "moving" in opposite directions, so that when one is increasing the other is decreasing and vice versa.

  Physically, this corresponds to the force $F = -kx$ being a linear restoring force and therefore both opposite and proportional to the displacement: when x is at its maximal positive value $x = +A$, then the acceleration $a = F/m = -kx/m = -\omega^2 x$ is at its "maximal" negative value[2] $a = -\omega^2 A$, while at the equilibrium point $x = 0$ there is no spring force and therefore no acceleration. Etc., etc.

- The velocity v is a quarter cycle out of phase with the location x, so that when x is at its extremal values $x = \pm A$, the velocity vanishes, while when $x = 0$ the velocity is at one of its extremal values $v = \pm\omega A$.

  Physically this corresponds to the points $x = \pm A$ being the endpoints of the motion: at these turn-around points the mass is for an instant at rest. And when the mass passes through the equilibrium point $x = 0$, it has just been returned, starting from rest at an endpoint, by a spring force that has been

[2] Yeah, yeah, yeah, really we should say *minimal value*, but we're just trying to be clear, all right? Give us a break.

pulling it always back toward the equilibrium point and that has therefore been continually speeding up the mass. Since the spring force will reverse direction and begin slowing the mass down after it passes the equilibrium point, the velocity has attained its extremal value at that point. In the first and third quarters of fig. (9.2), you can see the velocity building from $v = 0$ to $v = \pm\omega A$ as the mass travels from an endpoint to the equilibrium point. In the second and forth quarters of fig. (9.2), the mass is traveling from the equilibrium point to an endpoint, and you can see the reverse happening: the velocity dies away from $v = \pm\omega A$ to $v = 0$.

- Finally, bear in mind that fig. (9.2) is plotting the location as a function of *time*, that is, $x = x(t)$ and not something like $y = y(x)$; the sinusoidal curve shown for $x(t)$ in fig. (9.2) is *not* the path followed by the oscillating mass: the oscillation is one-dimensional, back and forth along a straight line, as was shown in fig. (9.1).

  It being customary in introductory texts, in order to capture the interest of new generations and motivate them to pursue the subject further, to crush your imagination into dust and rob you of the will to live, fig. (9.3) shows explicitly the relation between the actual motion and the plots of x, v, and a as functions of time. In the bottom of the figure, the horizontal time axis is aligned with that of the plots of the x, v, and a curves, and the vertical

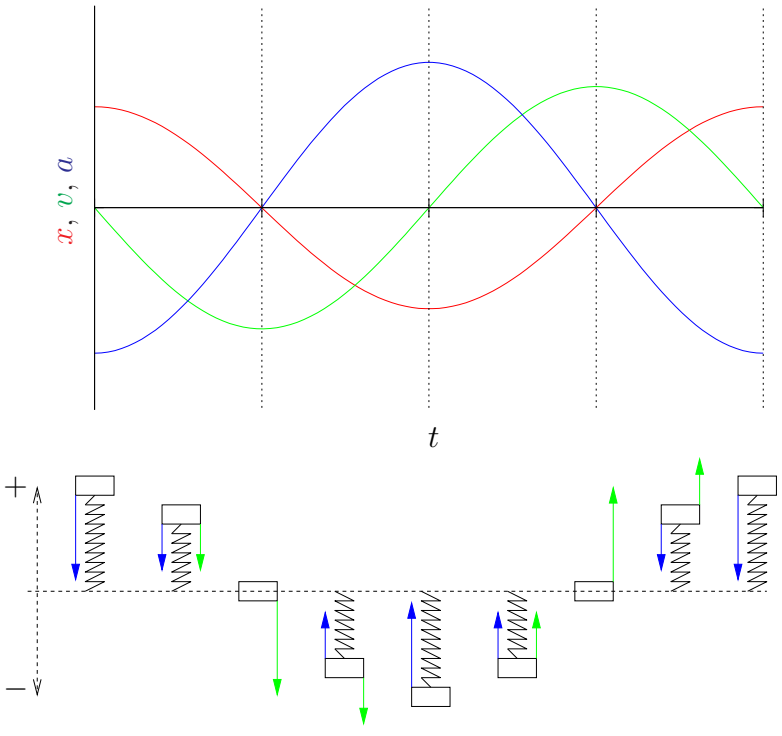


Figure 9.3: Must . . . Remain . . . Conscious . . . Aarrgh . . .

axis is the mass's location x, with the positive and negative directions as labeled. The "coils" represent how far the spring has been stretched or compressed from its equilibrium position. The green and blue arrows indicate the directions and, by their lengths, the relative magnitudes of the mass's velocity and acceleration, respectively. (These arrows have been drawn on the left and right sides of the mass simply for visibility. Also, for æsthetic reasons there is no correlation between the velocity and acceleration scales of the v and a curves and those of the arrows.)

We discussed on p.268 why harmonic motion like that of a mass on the end of a spring is so important. Among the reasons was that a great variety of physical systems experience harmonic oscillations under very commonly occurring conditions. In particular, as we will show in §9.7, any physical system that is in a stable equilibrium will, when perturbed slightly, oscillate to a first approximation harmonically about that equilibrium. These kinds of oscillations are known technically as small oscillations or small vibrations. For example,

- Anything reasonably rigid that is "just sitting there" — a bowl, a piece of furniture, whatever — is in a sort of stable equilibrium and will oscillate harmonically if you give it a whack. Often the resulting vibrations are too small to be noticeable and are very quickly damped out by friction, but they are quite substantial by design in rocking chairs and, though not visible, can be heard when you "ping" the rim of a bowl. In a moderate earthquake, buildings will oscillate harmonically.[3]
- Piano wires, guitar and violin strings, drum skins, etc., are all held in a kind of equilibrium by the tension they are put under. When plucked or struck, they undergo very audible and sometimes even visible harmonic vibrations.
- The surface of water, held smooth and level by gravity, is in a similar sort of equilibrium. Though they become distorted in shallows and break on the beach, in deeper water the waves that the wind stirs up on lakes and the ocean are harmonic: their shape is sinusoidal.[4]
- As we will see in §9.4, as long as the swings aren't too big, pendula oscillate harmonically.

[3] More precisely, they will respond like a complex, three-dimensional system of coupled springs.

[4] The waves you generate by tossing a pebble into the water, because they are circular rather than straight, are somewhat more complex mathematically. Though they oscillate harmonically, their shape is described by Bessel functions rather than sine and cosine waves. Bessel functions also describe shape of the vibrations of drum skins.

9.2 Once Again, with Feeling

In this section we will look into solutions to the harmonic equation for a mass on an ideal spring with a bit more mathematical sophistication, including fitting solutions to boundary conditions.

The equation of motion for a mass on the end of a spring is

$$-kx = m\frac{d^2x}{dt^2} \tag{9.5}$$

In the previous section, we saw that one way to write the solution to this equation is, as in eq. (9.2), as a cosine oscillation with a phase φ:

$$x = A\cos(\omega t + \varphi) \tag{9.6}$$

where $\omega = \sqrt{k/m}$ and A is the amplitude of the oscillation. Although it is conventional to use the cosine in this relation, we could equally well have used the sine, since cosine and sine have the same shape and differ only in phase. In fact, it is also possible to write the solution as a superposition of sine and cosine oscillations: [5]

$$x = A'\cos\omega t + B'\sin\omega t \tag{9.7}$$

where ω is again $\sqrt{k/m}$ and the A' and B' are parameters that are equivalent to (though of course different from) the parameters A and φ of eq. (9.6).

There are two ways to see that eq. (9.7) satisfies the equation of motion (9.5). One way is to show that it is equivalent to solution (9.6):

$$\begin{aligned} A\cos(\omega t + \varphi) &= A(\cos\omega t\cos\varphi - \sin\omega t\sin\varphi) \\ &= (A\cos\varphi)\cos\omega t + (-A\sin\varphi)\sin\omega t \end{aligned}$$

which is indeed the same as eq. (9.7) if we take

$$A' = A\cos\varphi \qquad\qquad B' = -A\sin\varphi$$

The other way is to again verify by substitution that the solution works, by plugging the x of eq. (9.7) into the equation of motion (9.5). We leave this as an exercise for the reader.

A third way to write the solution to eq. (9.5) is in terms of complex exponentials: [6]

$$x = \tilde{A}e^{i\omega t} + \tilde{B}e^{-i\omega t} \tag{9.8}$$

[5] We have put primes on the A and B here in order to distinguish the A of eq. (9.7) from that of eq. (9.6); otherwise these primes would be omitted.

[6] Again, we have put twiddles (tildes) on the A and B here in order to distinguish them from the A's and B of eqq. (9.6) and (9.7); otherwise these twiddles would be omitted.

If you recall the Euler relation

$$e^{i\theta} = \cos\theta + i\sin\theta$$

then you can see that the $e^{i\omega t}$ and $e^{-i\omega t}$ of eq. (9.8) are, so to speak, just sine and cosine waves rolled together. Eq. (9.8) is the grown-up way to write the solution for x. The advantage of complex exponentials is that they are, like all exponential functions, their own derivatives and integrals (unlike the sine and cosine, the derivatives and integrals of which alternate with each other and have sign changes to boot). The complication is that complex exponentials have, of course, both real and imaginary parts, so that the parameters $\tilde{A}$ and $\tilde{B}$ are now complex rather than just real numbers. This means that where as there were two "degrees of freedom" in solutions (9.6) and (9.7) (A and φ, or A' and B'), there are at first blush four degrees of freedom in solution (9.8) — a real and an imaginary part for each of $\tilde{A}$ and $\tilde{B}$. But these four degrees of freedom are reduced to two by the constraint that x, which represents a displacement, must be a real number: if we explicitly break $\tilde{A}$ and $\tilde{B}$ into their real and imaginary parts by

$$\tilde{A} = \tilde{A}_{re} + i\tilde{A}_{im} \qquad\qquad \tilde{B} = \tilde{B}_{re} + i\tilde{B}_{im}$$

we have

$$\begin{aligned} x &= \tilde{A}e^{i\omega t} + \tilde{B}e^{-i\omega t} \\ &= (\tilde{A}_{re} + i\tilde{A}_{im})(\cos\omega t + i\sin\omega t) + (\tilde{B}_{re} + i\tilde{B}_{im})(\cos\omega t - i\sin\omega t) \\ &= \Big((\tilde{A}_{re} + \tilde{B}_{re})\cos\omega t + (-\tilde{A}_{im} + \tilde{B}_{im})\sin\omega t\Big) \\ &\qquad + i\Big((\tilde{A}_{im} + \tilde{B}_{im})\cos\omega t + (\tilde{A}_{re} - \tilde{B}_{re})\sin\omega t\Big) \end{aligned}$$

Since the left-hand side of this relation is real, the imaginary part of the right-hand side must vanish *for all times t*. Since the sine and cosine are functionally independent, this means that we must have both

$$\tilde{A}_{im} + \tilde{B}_{im} = 0 \qquad \text{and} \qquad \tilde{A}_{re} - \tilde{B}_{re} = 0$$

or in other words

$$\tilde{A}_{im} = -\tilde{B}_{im} \qquad \text{and} \qquad \tilde{A}_{re} = \tilde{B}_{re}$$

or in yet other words[7]

$$\tilde{A} = \tilde{B}^*$$

where the star denotes complex conjugation:

[7] If you are good with complex numbers, the quick way to see this is to note that since the right-hand side of eq. (9.8), like its left-hand side, must be real, the right-hand side must be equal to its own complex conjugate:

$$\tilde{A}e^{i\omega t} + \tilde{B}e^{-i\omega t} = \big(\tilde{A}e^{i\omega t} + \tilde{B}e^{-i\omega t}\big)^* = \tilde{A}^*e^{-i\omega t} + \tilde{B}^*e^{i\omega t}$$

which, since $e^{i\omega t}$ and $e^{-i\omega t}$ are functionally independent, means that $\tilde{A} = \tilde{B}^*$ and (what amounts to the same thing) $\tilde{A}^* = \tilde{B}$.

$$(x + iy)^* = x - iy$$

Thus the four apparent degrees of freedom in $\tilde{A}$ and $\tilde{B}$ are reduced to two: the constraint that x must be real gives two relations among the total of four real and imaginary parts of $\tilde{A}$ and $\tilde{B}$.

Note that no matter how you write the solution to the equation of motion (9.5), since this equation of motion is a second-order differential equation, there are two arbitrary parameters in its solutions (A and φ, or A' and B', or the two remaining degrees of freedom in $\tilde{A}$ and $\tilde{B}$). As we will see in §20.1, taking what is known as a Fourier transform of a differential equation of order n reduces it to an algebraic equation of order n. The n roots of this algebraic equation correspond to n independent functions that satisfy the differential equation in any linear combination, and the coefficients of this linear combination constitute the n arbitrary constants (parameters) of the solution to the differential equation.

Physically, the values of these two parameters are determined by boundary conditions that stipulate either a specific value of $x(t)$ (the location of the mass is at some particular time) or a specific value of $v(t) = \dot{x}(t)$ (the velocity of the mass at some particular time). Often these boundary conditions are initial conditions that specify the location x_0 and velocity $v_0 = dx/dt|_{t=0}$ at time $t = 0$.

For example, suppose that we want to fit solution (9.6) to the conditions that the mass is at location x_0 and moving with velocity v_0 at time $t = 0$. Setting $x = x_0$ and $t = 0$ in eq. (9.6), we have

$$x_0 = A\cos\varphi \tag{9.9}$$

Taking a derivative of eq. (9.6) to obtain a result for velocity gives

$$v = \frac{dx}{dt} = -\omega A\sin(\omega t + \varphi)$$

which, if we set $v = v_0$ and $t = 0$, becomes

$$v_0 = -\omega A\sin\varphi \tag{9.10}$$

The easiest way to solve these equations for the parameters A and φ is to first take the ratio of eq. (9.10) to eq. (9.9):

$$-\frac{v_0}{\omega x_0} = \tan\varphi$$

which yields

$$\varphi = -\tan^{-1}\left(\frac{v_0}{\omega x_0}\right)$$

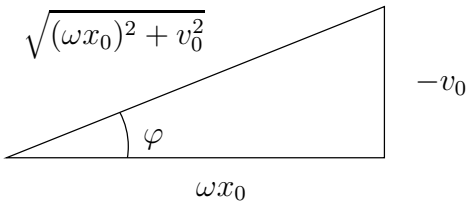


Figure 9.4: Ye Olde Triangle Trick

By the old triangle trig trick (illustrated in fig. (9.4)), we thus have

$$\cos\varphi = \frac{\omega x_0}{\sqrt{(\omega x_0)^2 + v_0^2}}$$

so that from eq. (9.9) we obtain

$$A = \frac{x_0}{\cos\varphi} = \frac{1}{\omega}\sqrt{(\omega x_0)^2 + v_0^2} = \sqrt{x_0^2 + \left(\frac{v_0}{\omega}\right)^2}$$

9.3 Some Practical Considerations

When working with masses on springs, it is not infrequently convenient to express the total energy

$$E = K + U = \tfrac{1}{2}mv^2 + \tfrac{1}{2}kx^2$$

in terms of its values at the equilibrium and endpoints. At the equilibrium point $x = 0$, the potential energy vanishes, and the speed and kinetic energy are maximal. Thus

$$E = \tfrac{1}{2}mv_{max}^2$$

At the endpoints (the turn-around points) $x = \pm A$, the mass is for an instant at rest, so that the kinetic energy vanishes and the potential energy is at its maximal value $\frac{1}{2}k(\pm A)^2 = \frac{1}{2}kA^2$. Thus

$$E = \tfrac{1}{2}kA^2$$

Not a big deal, but using one or the other of the above expressions for E is sometimes handy when solving problems.

When dealing with masses on vertical springs, the linearity of the spring force makes it possible to work in terms of a new equilibrium point that implicitly takes into account the weight of the mass.

Suppose we have a mass m on a vertical spring of spring constant k. The equilibrium point of the spring is still, of course, at $x = 0$, but the weight

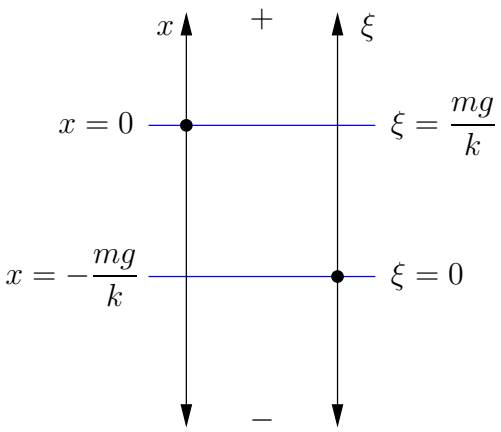


Figure 9.5: The Relation Between x and ξ

of the mass must be included in the equation of motion: if we make up the positive direction, $F = ma$ becomes

$$-kx - mg = m\frac{d^2x}{dt^2}$$

The spring force balances the weight of the mass at $-kx - mg = 0$, that is, at $x = -mg/k$, and we can make this the origin of a new coordinate ξ, as shown in fig. (9.5). The relation between x and ξ is therefore

$$\xi = x + \frac{mg}{k}$$

or, equivalently,

$$x = \xi - \frac{mg}{k}$$

Using this to rewrite the equation of motion in terms of ξ rather than x, we have

$$\begin{aligned} -kx - mg &= m\frac{d^2x}{dt^2} \\ -k\left(\xi - \frac{mg}{k}\right) - mg &= m\frac{d^2}{dt^2}\left(\xi - \frac{mg}{k}\right) \\ -k\xi &= m\frac{d^2\xi}{dt^2} \end{aligned}$$

That is, the equation of motion for the vertical spring is identical to that for a horizontal spring if we work in terms of the "shifted" equilibrium point that takes into account the weight of the mass. We will therefore get harmonic oscillations of frequency $\omega = \sqrt{k/m}$ centered about this new equilibrium point. Note that it was critical in this derivation that the spring force be

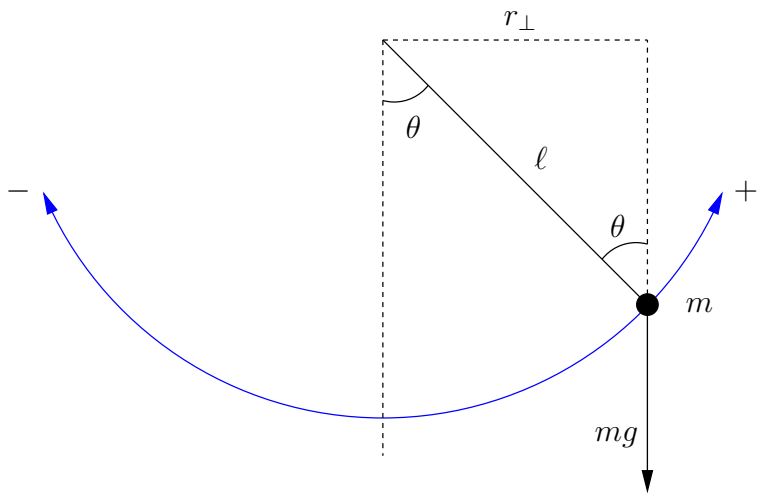


Figure 9.6: A Simple Pendulum

linear: were the spring force something like $-kx^3$ rather than $-kx$, then the left-hand side of the above equation of motion would not have worked out to simply $-k\xi$.

9.4 Pendula

While the harmonic motion of a mass on the end of a spring is translational, that of a pendulum is rotational: it rotates back and forth about a fixed pivot point, driven by the torque due to gravity. As you can see from fig. (9.6), the torque arm for the weight mg of the bob is $r_\perp = \ell \sin\theta$. If we use the sign convention shown in fig. (9.6) for the angular directions, at the moment illustrated θ and hence $\sin\theta$ are positive, but the torque due to gravity is in the negative direction. Thus

$$\tau_{grav} = -(\ell \sin\theta)(mg) = -mg\ell \sin\theta$$

Our rotational equation $\tau = I\alpha$ therefore becomes

$$-mg\ell \sin\theta = I\alpha = I\frac{d^2\theta}{dt^2} \tag{9.11}$$

where we have used $\alpha = d^2\theta/dt^2$. This differential equation for $\theta(t)$ is much more involved than the one for a mass on the end of a spring; the exact solution is an elliptic integral, which we really don't want to get into. Fortunately, we don't have to: as long as the swings of the pendulum don't go too far out to either side, θ will be small, and when θ is small, $\sin\theta \approx \theta$. Eq. (9.11) then reduces to

$$-mg\ell\theta \approx I\frac{d^2\theta}{dt^2}$$

which is identical to eq. (9.1),

$$-kx = m\frac{d^2x}{dt^2}$$

if we make the following correspondences:

Pendulum		Spring
θ	$\leftrightarrow$	x
$mg\ell$	$\leftrightarrow$	k
I	$\leftrightarrow$	m

For small swings, the oscillations of a pendulum are therefore approximately harmonic, with (by analogy to $\omega = \sqrt{k/m}$ from eq. (9.2)) angular frequency [8]

$$\omega = \sqrt{\frac{mg\ell}{I}} \tag{9.12}$$

How good an approximation was setting $\sin\theta \approx \theta$? The largest contribution to the deviation from this approximation will come from the next term in the Taylor expansion of $\sin\theta$:

$$\sin\theta \approx \theta - \frac{1}{3!}\theta^3$$

If the approximation is to be good to about, say, 1%, we might naively expect that we would therefore need

$$\frac{\frac{1}{3!}\theta^3}{\theta} = \frac{1}{6}\theta^2 \lesssim 0.01$$

or $\theta \lesssim 0.24$ (about 14°). But even without knowing anything about the elliptic integrals involved in the exact solution of the pendulum motion we can do better than this naive estimate. We used the approximation $\sin\theta \approx \theta$ solely for the gravitational torque driving the pendulum, and this torque most directly determines the angular acceleration, which, in terms of physical dimensions, is inverse time squared:

$$\tau \sim \alpha \sim \frac{1}{t^2}$$

Since $\sin\theta$ is actually less than θ, using $\sin\theta \approx \theta$ overestimates the torque and angular acceleration and therefore underestimates the timing and period of

[8] Our notation at this juncture is unavoidably bad: above we have referred to θ and the corresponding angular acceleration $\alpha = \ddot{\theta}$. But the ω to which we refer in eq. (9.12) is *not* $\dot{\theta}$ (which would be the angular velocity of the bob and would vary from zero at the endpoints of the swing to some maximal value at the bottom) but rather the angular frequency of the pendulum oscillations — apples and oranges.

the pendulum: the actual period of the pendulum will be longer. If we want the *period* of the pendulum to be accurate to 1%, then, since $(1.01)^2 \approx 1.02$, we only need for the torque and our approximation $\sin\theta \approx \theta$ to be accurate to 2% and hence[9]

$$\frac{\frac{1}{3!}\theta^3}{\theta} = \frac{1}{6}\theta^2 \lesssim 0.02$$

which yields $\theta \lesssim 0.35$ (about 20°).

This reasoning from physical dimensions may seem a bit flimsy, but, while it certainly isn't rigorous, it falls in the category of "good enough." Still not rigorous, but less flimsy, would be starting from our result eq. (9.12) to obtain for the harmonic approximation of the period

$$T = \frac{2\pi}{\omega} = 2\pi\sqrt{\frac{I}{mg\ell}}$$

Since the approximation $\sin\theta \approx \theta$ was to the torque and hence to the $mg\ell$ part of this result, we have $T \sim 1/\sqrt{mg\ell}$ or, equivalently,

$$\frac{1}{T^2} \sim mg\ell$$

By the same reasoning as above this leads to the conclusion that the period T will be accurate to 1% when our approximation for $mg\ell$ and the torque, and therefore $\sin\theta \approx \theta$, is accurate to 2%.

Were we to delve into the exact solution for the pendulum motion and elliptic integrals of the first kind, we would find that for the period to be accurate to 1% in fact requires $\theta \lesssim 0.4$ (about 23°). For larger swings, the approximation becomes progressively worse, and for $\theta = \frac{\pi}{2}$, approximating the value of $\sin\frac{\pi}{2} = 1$ by $\frac{\pi}{2}$ would be awful even by the decadent standards of contemporary society. To see the deviation of the actual pendulum motion from our harmonic approximation for larger swings, we can numerically integrate the exact pendulum eq. (9.11),

$$\frac{d^2\theta}{dt^2} = -\frac{mg\ell}{I}\sin\theta \tag{9.13}$$

To do such a numerical integration, we need to give numerical values to any parameters in the equation being integrated — in this case, to the m, g, ℓ,

[9] When performing approximations, it is often very handy to remember that if $|\epsilon| \ll 1$, then, from the leading two terms in the Taylor expansion, $(1\pm\epsilon)^n \approx 1 \pm n\epsilon$. In particular, frequently useful are

$$\frac{1}{1\pm\epsilon} \approx 1 \mp \epsilon \qquad\qquad \sqrt{1\pm\epsilon} \approx 1 \pm \tfrac{1}{2}\epsilon$$

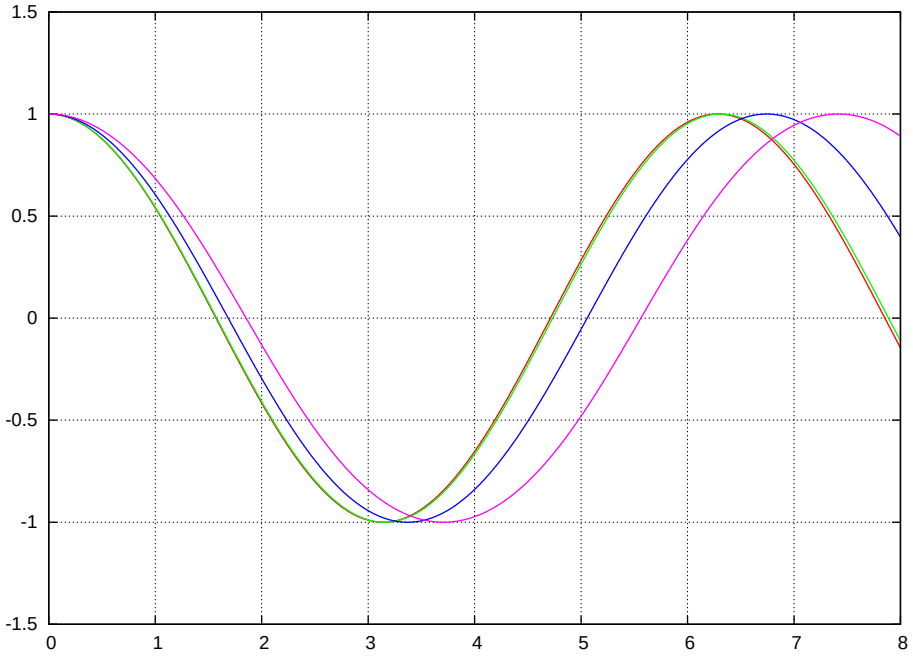


Figure 9.7: It's Alive!

and I. Choosing specific numerical values for these parameters would, of course, result in a loss of generality we would like to avoid if possible, and it is possible: if we note that, according to eq. (9.12), $\sqrt{mg\ell/I}$ has inverse time dimensions, we can rewrite eq. (9.13) in terms of a dimensionless time parameter

$$\tilde{t} = t\sqrt{\frac{mg\ell}{I}}$$

Thus

$$\frac{d^2\theta}{dt^2} = \frac{d^2\theta}{d\left(\tilde{t}\sqrt{\frac{I}{mg\ell}}\right)^2} = \frac{mg\ell}{I}\frac{d^2\theta}{d\tilde{t}^2}$$

so that eq. (9.13) simplifies to

$$\frac{d^2\theta}{d\tilde{t}^2} = -\sin\theta \tag{9.14}$$

Eq. (9.14) can be integrated numerically without arbitrarily assigning numerical values to any parameters. Fig. (9.7) shows the results of such a numerical integration with the pendulum initially (that is, at $\tilde{t} = 0$) at the positive endpoint of its swing. The green, blue, and magenta curves in fig. (9.7) correspond to the cases where the amplitude of the swing is $\frac{\pi}{12}$, $\frac{\pi}{3}$, and $\frac{\pi}{2}$,

respectively (15°, 60°, and 90°). With the approximation $\sin\theta \approx \theta$, the solution would have been $\theta = \cos\tilde{t}$, which, for comparison, is plotted in red in fig. (9.7). To make the comparison easier, all of the curves have been normalized along the vertical axis so that they oscillate between -1 and 1.[10] As you can see, the larger the amplitude, the longer the period. Physically, this is what we expect: since the torque that is driving the oscillations is proportional to $\sin\theta$, and since $\sin\theta$ falls progressively shorter of the θ required for harmonic motion of constant period as θ increases, the outer regions of wider swings are driven by progressively weaker torques and therefore the time taken by each swing progressively increases. For an amplitude of $\frac{\pi}{12}$, the deviation from harmonicity is, as we expected, very slight, with the result that the green curve almost overlaps the red curve. For amplitudes of $\frac{\pi}{3}$ and $\frac{\pi}{2}$, the deviation is very significant. What is not as clearly visible from fig. (9.7) is that it is not simply that the period is getting longer as the amplitude increases, but also that the curve is becoming less harmonic, that is, is deviating more and more from the sinusoidal shape that characterizes harmonic motion.

Fig. (9.6) showed what is known as an idealized simple pendulum: a point-mass bob at the end of a massless string or rod. Since $I = mr^2$ for a point mass, the bob's moment of inertia is $I = m\ell^2$, and eq. (9.12) reduces to

$$\omega = \sqrt{\frac{mg\ell}{m\ell^2}} = \sqrt{\frac{g}{\ell}} \tag{9.15}$$

Eq. (9.12) is, of course, much more general: as long as I is the moment of inertia of the pendulum about the pivot point and ℓ the distance from the pivot to the center of mass of the pendulum (where gravity effectively acts), then eq. (9.12) holds for any kind of pendulum. Pendula with mass distributions more complex than that of a simple pendulum are known as compound pendula. For example, a pendulum in which the rod had a nonzero mass, or in which the bob was a disk of finite radius rather than a point mass, would constitute a compound pendulum.

It is also possible to analyze the motion of the pendulum by energy. Since the pendulum is undergoing a pure rotation about the pivot point, the expression for its kinetic energy (in the more general case of the compound

[10] As opposed to between $-\frac{\pi}{12}$ and $\frac{\pi}{12}$ for the swing of amplitude $\frac{\pi}{12}$, $-\frac{\pi}{3}$ and $\frac{\pi}{3}$ for the swing of amplitude $\frac{\pi}{3}$, and (you guessed it) $-\frac{\pi}{2}$ and $\frac{\pi}{2}$ for the swing of amplitude $\frac{\pi}{2}$.

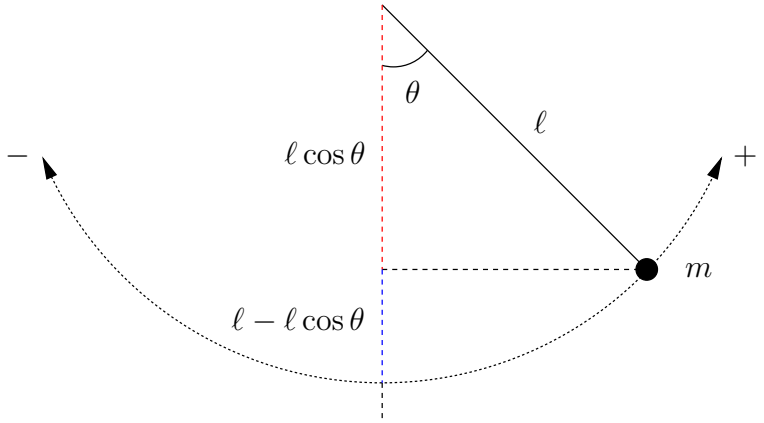


Figure 9.8: A Simple Pendulum

pendulum) is [11]

$$K = \tfrac{1}{2}I\omega^2 = \tfrac{1}{2}I\left(\frac{d\theta}{dt}\right)^2 \tag{9.16}$$

The gravitational potential energy of the pendulum will depend on the location of the pendulum's center of mass. As you can see from fig. (9.8), the height of the pendulum above the lowest point in its swing is

$$h = \ell(1 - \cos\theta)$$

so that potential energy of the pendulum is

$$U = mg\ell(1 - \cos\theta)$$

Although fig. (9.8) shows only a simple pendulum, this result will hold for the compound pendulum as well, with ℓ in the case of the compound pendulum taken to be the distance from the pivot to the center of mass. For small oscillations (that is, for small θ),

$$\cos\theta \approx 1 - \tfrac{1}{2}\theta^2$$

and the potential energy becomes

$$U \approx mg\ell\Big(1 - (1 - \tfrac{1}{2}\theta^2)\Big) = \tfrac{1}{2}mg\ell\theta^2 \tag{9.17}$$

Now, if we had a mass m oscillating on the end of a spring of spring constant k, the expressions for the kinetic and potential energies would be

$$K = \tfrac{1}{2}mv^2 = \tfrac{1}{2}m\left(\frac{dx}{dt}\right)^2 \qquad\qquad U = \tfrac{1}{2}kx^2$$

[11] Be careful here not to confuse the instantaneous angular velocity $\omega = d\theta/dt$ of the pendulum with the angular frequency of its oscillations: these are two totally different animals. It is the latter that is given by eqq. (9.15) and (9.12) but of course the former that appears in eq. (9.16).

Comparing these expressions for the kinetic and potential energies to those of the pendulum (eqq. (9.16) and (9.17)), we see that they are really the same if we make the following correspondences:

Spring		Pendulum
x	$\leftrightarrow$	θ
m	$\leftrightarrow$	I
k	$\leftrightarrow$	$mg\ell$

The motion of the pendulum will therefore be the same as that of a mass on the end of a spring: just as the x coordinate of the mass on a spring oscillates harmonically at angular frequency $\omega = \sqrt{k/m}$, the angular displacement θ of a pendulum will oscillate harmonically at angular frequency

$$\omega = \sqrt{\frac{mg\ell}{I}}$$

which is just eq. (9.12).

The simple-pendulum result (9.15), $\omega = \sqrt{g/\ell}$, makes it possible to obtain a fairly accurate result for g quickly and easily: you just construct a simple pendulum by tying a weight to the end of a string, measure the length ℓ of the string, and use a stopwatch to get a result for the pendulum's period T (from which you can get a result for ω). If you time many swings and use reasonable care when measuring ℓ, you can easily get a three-digit result for g right in your own kitchen (or wherever it is at home that you do your scientific experiments). In fact, the field apparatus that the National Geologic Survey most commonly uses to measure g at various places around the country is a double pendulum, with bobs simultaneously swinging in opposite directions so that there is no error due to rocking.[12]

Just so you are aware of it, there is another sort of pendulum known as a torsional pendulum, which consists of a horizontal disk suspended from its center on a vertical wire. The pendulum is set going by rotating the disk a little and letting it go: as the wire alternately untwists and retwists itself, the disc undergoes an approximately harmonic oscillation, rotating back and forth sort of like the mechanism you may have seen in some clocks.

[12] For more accurate measurements, there is another apparatus that uses lasers to time free-fall in a vacuum. Apparently this apparatus is about the size of truck, though, so it isn't very portable.

9.5 Damped Harmonic Oscillations

We will assume that the damping force is proportional to the velocity,

$$F_{fric} = -bv$$

where $b > 0$ is the constant of proportionality and the negative sign indicates that this frictional force is of course always opposite in direction to the velocity v. The equation of motion of our damped spring oscillation is then

$$\begin{aligned} F &= ma \\ -kx - bv &= ma \\ -kx - b\frac{dx}{dt} &= m\frac{d^2x}{dt^2} \end{aligned}$$

which, if we divide through by m and move all the terms to the right-hand side, becomes

$$0 = \frac{d^2x}{dt^2} + \frac{b}{m}\frac{dx}{dt} + \frac{k}{m}x \tag{9.18}$$

If we were able to do Fourier transforms, we could reduce this second-order differential equation to a quadratic equation and get the solution directly.[13] As the next best thing, we will instead test whether the trial solution $x = Ae^{i\omega t}$, where A and ω are constant parameters, will satisfy eq. (9.18):

$$\begin{aligned} 0 &= \frac{d^2}{dt^2}(Ae^{i\omega t}) + \frac{b}{m}\frac{d}{dt}(Ae^{i\omega t}) + \frac{k}{m}(Ae^{i\omega t}) \\ &= -A\omega^2 e^{i\omega t} + iA\omega\frac{b}{m}e^{i\omega t} + A\frac{k}{m}e^{i\omega t} \end{aligned}$$

Dividing through by $Ae^{i\omega t}$, we arrive at

$$0 = -\omega^2 + i\frac{b}{m}\omega + \frac{k}{m}$$

Thus $x = Ae^{i\omega t}$ is in fact a solution of the equation of motion if we take

$$\omega = i\frac{b}{2m} \pm \frac{1}{2}\sqrt{-\frac{b^2}{m^2} + 4\frac{k}{m}} = i\frac{b}{2m} \pm \sqrt{\frac{k}{m} - \frac{b^2}{4m^2}} \tag{9.19}$$

The second-order differential eq. (9.18) should have two and only two "root" solutions. The trial solution $x = Ae^{i\omega t}$ satisfies eq. (9.18) for two values of ω (the $\pm$ cases in eq. (9.19)), and these two solutions are therefore

[13] Fourier transforms are actually presented in §20.1, but introducing them here would take us too far afield.

the solutions to eq. (9.18) — there are no others. The general solution of eq. (9.18) is a linear combination of these two solutions:

$$\begin{aligned} x &= A\exp\left(i\left(i\frac{b}{2m}+\sqrt{\frac{k}{m}-\frac{b^2}{4m^2}}\right)t\right) \\ &\qquad + B\exp\left(i\left(i\frac{b}{2m}-\sqrt{\frac{k}{m}-\frac{b^2}{4m^2}}\right)t\right) \\ &= e^{-\frac{bt}{2m}}\left(A\exp\left(+i\sqrt{\frac{k}{m}-\frac{b^2}{4m^2}}\,t\right)\right. \\ &\qquad \left. +B\exp\left(-i\sqrt{\frac{k}{m}-\frac{b^2}{4m^2}}\,t\right)\right) \end{aligned}$$

If we define ω_0 and β by $\omega_0 = \sqrt{k/m}$ (the angular frequency of oscillation without damping) and $\beta = b/2m$, this becomes

$$x = e^{-\beta t}\left(Ae^{+i\sqrt{\omega_0^2-\beta^2}\,t} + Be^{-i\sqrt{\omega_0^2-\beta^2}\,t}\right) \tag{9.20}$$

The behavior of this solution depends principally on the whether $\omega_0^2 - \beta^2$ is positive, negative, or zero:

- When $\omega_0 > \beta$ (corresponding to a relatively small b, that is, to light damping), we see from the complex exponentials that we get an oscillation at angular frequency $\sqrt{\omega_0^2 - \beta^2}$. This oscillation is slower than the oscillation frequency ω_0 that we would have without damping, and the amplitude of the oscillation dies away with the damping factor $e^{-\beta t}$ outside the parentheses. This damped oscillation is plotted in red in fig. (9.9), along with its envelope $\pm e^{-\beta t}$ in blue.

- When $\omega_0 < \beta$ (corresponding to relatively large b, that is, to over-damping), $\sqrt{\omega_0^2 - \beta^2}$ is imaginary, with the result that the exponentials inside the parentheses are real rather than complex. There is thus no oscillation at all — just a damped return to equilibrium in the form of a negative exponential.

- When $\omega_0 = \beta$ (the critical case), the exponents inside the parentheses vanish, so that $x \propto e^{-\beta t}$: like the over-damped case, there is no oscillation, but the return to equilibrium is more rapid than in the case of over-damping.

Real examples of damped oscillations are ubiquitous: all of the myriad forms of harmonic motion that occur in everyday contexts involve at least

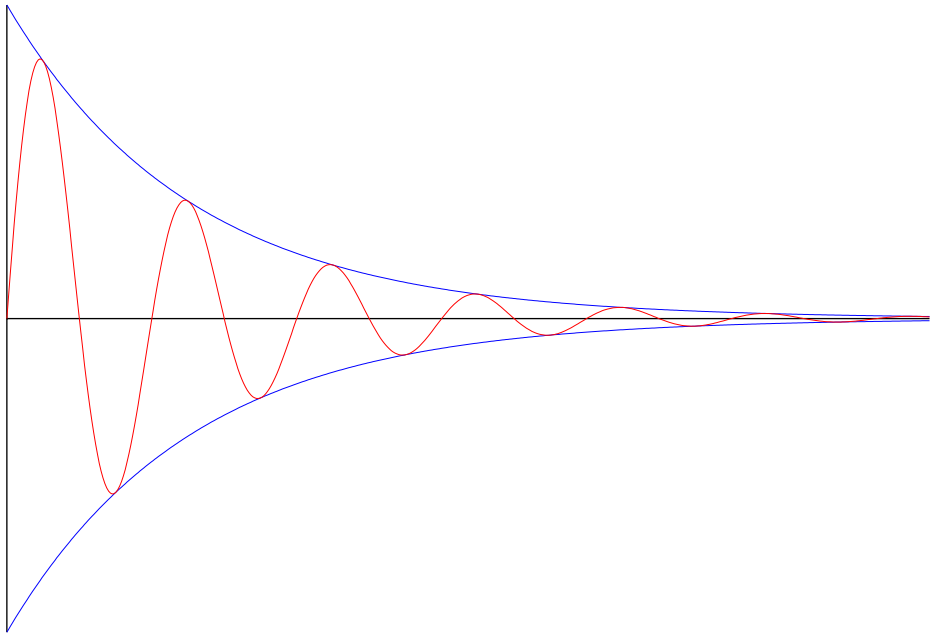


Figure 9.9: Light Damping

some friction that damps out the oscillation and eventually brings everything to rest. One common example is a swinging door: if the door is well designed, its mechanism should be critically damped so that it ceases swinging as quickly as possible.

9.6 Driven Damped Harmonic Oscillations

Suppose now that the oscillator is driven by an applied harmonic force of angular frequency Ω,

$$F_{applied} = F_0 \cos \Omega t = \tfrac{1}{2} F_0 \left(e^{i\Omega t} + e^{-i\Omega t} \right)$$

where F_0 is a positive constant. With the addition of this driving force to the F side of $F = ma$, the equation of motion (9.18) now becomes

$$-kx - b\frac{dx}{dt} + \tfrac{1}{2} F_0 e^{i\Omega t} = m\frac{d^2x}{dt^2}$$

where we have, even though it may seem very odd, for now included just the $e^{+i\Omega t}$ part of $F_{applied}$ for simplicity — we can easily put the $e^{-i\Omega t}$ part back in later, and dealing first with just the $e^{+i\Omega t}$ part will make the calculations a lot easier. If we isolate the F_0 term, divide through by m, and again define

$\omega_0 = \sqrt{k/m}$ and $\beta = b/2m$, we have

$$\begin{aligned} \frac{F_0}{2m} e^{i\Omega t} &= \frac{d^2x}{dt^2} + \frac{b}{m}\frac{dx}{dt} + \frac{k}{m}x \\ &= \frac{d^2x}{dt^2} + 2\beta\frac{dx}{dt} + \omega_0^2 x \end{aligned} \tag{9.21}$$

We again test a trial solution $x = Ae^{i\omega t}$. Plugging this trial solution into the equation of motion (9.21) and pulling out common factors, we have

$$\frac{F_0}{2m} e^{i\Omega t} = A\Big(-\omega^2 + 2i\beta\omega + \omega_0^2\Big)\, e^{i\omega t} \tag{9.22}$$

Since $e^{i\omega t}$ and $e^{i\Omega t}$ are otherwise functionally independent, in order for this equation to hold for all times t we must have $\omega = \Omega$. When $\omega = \Omega$, eq. (9.22) yields

$$A = \frac{F_0/2m}{\omega_0^2 - \Omega^2 + 2i\beta\Omega}$$

and thus

$$x = Ae^{i\omega t} = \frac{F_0/2m}{\omega_0^2 - \Omega^2 + 2i\beta\Omega} e^{i\Omega t}$$

We now have a solution for x — at least for the $e^{+i\Omega t}$ part of the driving force — , but to make physical sense of this solution we need to simplify it. It will be much easier to interpret our result if we write the complex factor

$$\frac{1}{\omega_0^2 - \Omega^2 + 2i\beta\Omega}$$

in polar form. To accomplish this, we first use the usual trick to make the denominator real and get the factor into the more conventional form $x + iy$:

$$\begin{aligned} \frac{1}{\omega_0^2 - \Omega^2 + 2i\beta\Omega} &= \frac{1}{\omega_0^2 - \Omega^2 + 2i\beta\Omega}\left(\frac{\omega_0^2 - \Omega^2 - 2i\beta\Omega}{\omega_0^2 - \Omega^2 - 2i\beta\Omega}\right) \\ &= \frac{\omega_0^2 - \Omega^2 - 2i\beta\Omega}{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2} \\ &= \frac{\omega_0^2 - \Omega^2}{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2} + i\,\frac{-2\beta\Omega}{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2} \end{aligned} \tag{9.23}$$

Now recall that the complex number

$$z = x + iy$$

can be written in the polar form

$$z = \rho\, e^{i\varphi}$$

where

$$\rho = \sqrt{x^2 + y^2} \qquad x = \rho\cos\varphi$$
$$\varphi = \tan^{-1}\left(\frac{y}{x}\right) \qquad y = \rho\sin\varphi$$

For the factor (9.23), ρ is

$$\rho = \sqrt{\left(\frac{\omega_0^2 - \Omega^2}{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}\right)^2 + \left(\frac{-2\beta\Omega}{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}\right)^2}$$
$$= \frac{1}{\sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}}$$

and φ is

$$\varphi = \tan^{-1}\left(\frac{\dfrac{-2\beta\Omega}{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}}{\dfrac{\omega_0^2 - \Omega^2}{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}}\right) = \tan^{-1}\left(\frac{-2\beta\Omega}{\omega_0^2 - \Omega^2}\right)$$

Thus the factor (9.23) can be written as

$$\frac{1}{\omega_0^2 - \Omega^2 + 2i\beta\Omega} = \frac{1}{\sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}}\, e^{i\varphi}$$

and the trial solution x as

$$x = \frac{F_0/2m}{\sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}}\, e^{i(\Omega t - \varphi)}$$

where

$$\varphi = \tan^{-1}\left(\frac{-2\beta\Omega}{\omega_0^2 - \Omega^2}\right) \tag{9.24}$$

At this point we recall that this is the solution only for the $e^{i\Omega t}$ part of the driving force $F_{applied}$; to include the $e^{-i\Omega t}$ part of $F_{applied}$, we just add to x the same solution with $-\Omega$ in place of Ω:

$$x = \frac{F_0/2m}{\sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}}\left(e^{i(\Omega t - \varphi)} + e^{i(-\Omega t - \varphi)}\right)$$
$$= \frac{F_0/2m}{\sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}}\left(2\cos(\Omega t - \varphi)\right)$$
$$= \frac{F_0/m}{\sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}}\cos(\Omega t - \varphi) \tag{9.25}$$

In this simplified form, the behavior of our full solution (9.25) is more easily analyzed:

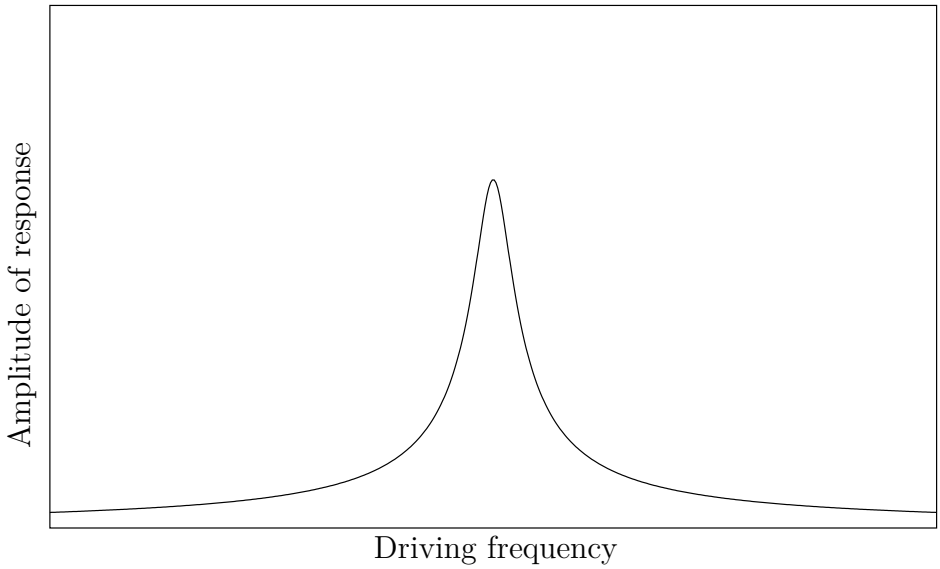


Figure 9.10: Resonance in a Damped, Driven Oscillator

- As you can see from the cosine factor, the response of the oscillator to the driving force is always a harmonic oscillation. Moreover, the oscillator always responds at the frequency Ω of the driving force, regardless of the natural frequency $\omega_0 = \sqrt{k/m}$ of the oscillator.

- The amplitude of the response is

$$\frac{F_0/m}{\sqrt{(\omega_0^2 - \Omega^2)^2 + 4\beta^2\Omega^2}}$$

  The effect of friction is in the β term, which, as you can see, limits the response of the oscillator: the more friction, the larger β and the smaller the amplitude of the oscillations. And as a function of the driving frequency Ω, this amplitude is, again depending on the value of β, a more or less sharply peaked function of the driving frequency Ω, as shown in fig. (9.10). The peak, known as the resonance, is at $\Omega = \omega_0$, which corresponds to the oscillator being driven at its natural frequency. In this case the amplitude reduces to

$$\frac{F_0}{2m\beta\omega_0} = \frac{F_0}{b\omega_0}$$

  which blows up as the constant of proportionality b for the damping force goes to zero. That is, the resonance is an infinite spike in the absence of damping. When $\Omega \neq \omega_0$, even in the absence of damping there is a limit to the excitation of the oscillator.

- Any solution to the nondriven equation of motion (9.18) may be superposed on solution (9.25). Since all such nondriven solutions to the damped oscillator die away because of damping, they are called transients.

Now let's make more intuitive sense of the solution. Unlike the case of a damped oscillator that is not driven, friction does not cause the oscillations of a driven damped oscillator to die away, because the driving force is continually pumping energy into the oscillator to compensate for the energy lost to friction; the effect of friction, as noted above, is merely to limit the amplitude of the oscillator's response, with that limit determined by the balance reached between the rates at which energy is supplied by the driving force and lost to friction. And the dependence of the response on the driving force can be understood in terms of pushing a friend on a swing: the closer the frequency of your pushes to the swing's natural frequency — that is, the closer your pushes are to being in sync with the swing's natural motion — , the bigger the resulting swings. When the frequency of your pushes is far from the swing's natural frequency, your pushes and the swing's own spring-like restoring force vary continuously from being completely in sync (in phase) to being completely out of sync (out of phase) with each other, so that your pushes are as likely to be working against the swing's motion as with it and the resulting swings are correspondingly small.

Real examples of driven, damped harmonic oscillations are again very common, though the driving forces are not usually harmonic. But since it turns out to be possible by what is known as Fourier analysis to resolve any driving force, no matter what its time dependence, into a linear combination of harmonic driving forces, the above results still apply quite generally. One familiar example is the response of your car's shock absorbers to holes and bumps in the road: when driving over a rough stretch of road, the ride will be rougher when the frequency at which you are hitting the holes and bumps is closer to the natural frequency of your shock absorbers.[14] In car commercials boasting of a smooth ride over a road made of logs or the like, all they have to do to look good is make sure the speed of the car is such that it hits the logs far from the natural frequency of the shock absorbers.[15]

Another "familiar" example is the shattering of crystal by an opera singer: if the singer sings at the pitch corresponding to the crystal's natural frequency — the pitch at which the crystal would ring if you pinged it with your finger

[14] If you are perverse, you might try to determine the natural frequency of your own car's shock absorbers by repeatedly driving over a rough stretch of road at various speeds: when the ride is roughest, the frequency at which you are hitting the holes and bumps matches this natural frequency.

[15] Our apologies if this has left you disillusioned about the honesty and integrity of capitalist enterprises.

— and the resonance peak is high enough, the glass may vibrate so intensely that it shatters.[16]

Driven oscillations and fig. (9.10) also help explain why the sky is blue. It turns out that the molecules of the atmosphere, being in stable bound states that constitute equilibria of sorts, to a good approximation respond like driven harmonic oscillators to the electric forces exerted on them by passing light waves (which consist of electromagnetic oscillations). For a more expansive explanation of why the sky is blue and why sunrises and sunsets are red, see §0.3.2.

9.7 Small Oscillations

Suppose a potential energy $U(x)$ has a stable equilibrium point at $x = x_e$. We can then do a Taylor expansion of U about the point x_e:

$$U(x) = U(x_e) + (x - x_e)\left[\frac{dU}{dx}\right]_{x=x_e} + \tfrac{1}{2}(x - x_e)^2\left[\frac{d^2U}{dx^2}\right]_{x=x_e} + \frac{1}{3!}(x - x_e)^3\left[\frac{d^3U}{dx^3}\right]_{x=x_e} + \cdots$$

Since it is only changes in potential energy that matter physically, we can trash the first term in this expansion.[17] The second term vanishes because x_e is an equilibrium point and thus an extremum (in this case, a minimum) of U, so that $dU/dx = 0$ at $x = x_e$. If now we restrict ourselves to small perturbations about x_e, so that the location x of the object doesn't stray too far from x_e, then $x - x_e$ is small, and to a good approximation we can neglect the $(x - x_e)^3$ and higher terms in the expansion. Thus we arrive at

$$U(x) \approx \tfrac{1}{2}(x - x_e)^2\left[\frac{d^2U}{dx^2}\right]_{x=x_e} = \tfrac{1}{2}\left[\frac{d^2U}{dx^2}\right]_{x=x_e}(x - x_e)^2$$

which, since $(x - x_e)^2$ is just the square of the displacement from the equilibrium point, looks just like the potential energy $\frac{1}{2}kx^2$ for a spring if we interpret $[d^2U/dx^2]_{x=x_e}$ as the effective spring constant.

[16] Personally, we think this whole business is urban legend; it seems highly doubtful that even the fattest opera singer could produce enough sound energy to shatter a glass. But in principle it's possible, and if it does occur this is the explanation for it.

[17] Alternatively, we can make x_e the point where $U = 0$ by setting $x_0 = x_e$ in

$$U(x) = -\int_{x_0}^{x} F\,dx$$

We thus arrive at a very important result: any sufficiently small perturbation about a stable equilibrium point will result in a harmonic oscillation identical to that of a mass on the end of a spring, with the effective spring constant being[18]

$$k_{\mathit{eff}} = \left.\frac{d^2U}{dx^2}\right|_{x=x_e} \tag{9.26}$$

For example, the angular frequency of the oscillation of a mass m about the equilibrium point would be given by

$$\omega = \sqrt{\frac{k_{\mathit{eff}}}{m}} = \sqrt{\frac{\left.\frac{d^2U}{dx^2}\right|_{x=x_e}}{m}}$$

When you use eq. (9.26) don't forget that the d^2U/dx^2 must be evaluated at the equilibrium point; this is a frequent cause of error that leads to all sorts of complete nonsense like spring "constants" that are functions of location.

Suppose, for example, that a small marble[19] of mass m is sitting in a bowl of which the height of the sides is given, in cylindrical coordinates, by

$$z = a\, e^{r/b}$$

where a and b are positive constants. The gravitational potential energy of the marble is then given by

$$U = mgz = mga\, e^{r/b}$$

If you give the marble a small ping at as it sits at the bottom of the bowl (that is, at $r = 0$), it will oscillate harmonically with

$$k_{\mathit{eff}} = \left[\frac{d^2}{dr^2}(mga\, e^{r/b})\right]_{r=0} = \left[\frac{mga}{b^2}\, e^{r/b}\right]_{r=0} = \frac{mga}{b^2}$$

and hence

$$\omega = \sqrt{\frac{k_{\mathit{eff}}}{m}} = \sqrt{\frac{ga}{b^2}}$$

One noteworthy feature of this result is that it is independent of the mass of the marble: marbles of all masses will oscillate at the same angular frequency.

[18] What if the second derivative of the potential energy U is zero? Minima and maxima of U would still be stable and unstable equilibrium points, respectively, and there would still be oscillations about the stable equilibrium points, but these oscillations would no longer be spring-like (that is, harmonic). Such oscillations are termed anharmonic.

[19] If you want to make the example more interesting, I suppose you could make it the Mutant Death Marble from Hell or something.

9.8 Wave Effects

9.8.1 Traveling Waves

The oscillations of a mass on a spring or a pendulum are stationary in the sense that the center of the oscillation is a stationary point: the mass and the pendulum are in continual motion, but they don't "go" anywhere. A traveling wave, on the other hand, not only oscillates but moves from place to place as it does so.

Perhaps the simplest example of a traveling wave is an ocean wave. If you place a buoy at a fixed location in the water, it will bob vertically up and down, reaching its highest point at each wave crest and its lowest at each wave trough.[20] This vertical motion turns out to be identical to the oscillation of a mass on the end of a spring. But at the same time, the water wave is also moving horizontally. This forward motion of the wave is known as propagation.

We won't get into the mathematical treatment of water and other traveling waves here, but we do want you to be aware that the vertical motion of the buoy is related to the forward motion of the water wave: in order for the oscillation and propagation to be consistent with each other, the period T of the oscillation must match up with the wavelength λ. That is, the time it takes the buoy to go from crest to trough and back to crest again must match up with the time it takes the wave to move forward the horizontal distance from one crest to the next. The speed v at which the wave is propagating horizontally must therefore be

$$v = \frac{\lambda}{T}$$

or, since the frequency f is the reciprocal of the period ($f = 1/T$),

$$v = \lambda f$$

For a water wave, the vertical direction of the oscillation is perpendicular to the horizontal direction of propagation. Traveling waves for which the oscillation and propagation are perpendicular to each other are known as transverse waves. Most traveling waves are transverse. Although they are, ironically, not directly visible, light and other electromagnetic waves are transverse waves; they consist of a simultaneous oscillation of electric and magnetic fields, with both of these oscillations being perpendicular to the direction of motion of the light wave.[21]

[20] By *fixed location* we mean that the buoy is not allowed to move horizontally. The actual motion of an object, such as one of those foam packing peanuts, would be approximately circular if it were floating freely on the surface.

[21] The direction along which the electric field oscillates is in fact the direction of polarization of the light.

Sound waves are an exception. Sound is a pressure oscillation in the air: all sound is produced by vibrating objects that, as they vibrate, alternately push against the air and pull back from it, which causes an oscillation in the air's density and pressure. This oscillatory pushing and pulling is along the same direction that the sound wave is moving, and sound waves are therefore termed longitudinal. The frequency of the sound wave determines its musical pitch.

9.8.2 Standing Waves

When you pluck a violin string, it turns out that you are sending a whole mess of sine waves of various frequencies and wavelengths down the length of the string in either direction. These waves are reflected back at the anchored ends of the string, with the result that the vibration of the string consists of numerous waves of various frequencies and wavelengths overlapped with each other and traveling in both directions. For most of the waves, this superposition will result in gibberish that will quickly be damped out. But if the wavelength of a wave is such that a whole number of half wavelengths match the length of the string (as shown for the case of $\frac{3}{2}$ of a wavelength in fig. (9.11)), it turns out that the waves of that wavelength reinforce each other as they bounce back and forth between anchor points. The result is a standing wave, a kind of resonance. The standing wave corresponding to fig. (9.11) is shown in fig. (9.12).

What you hear from a vibrating violin string is a superposition of the various standing waves, predominantly those of longer wavelength and lower frequency. Fig. (9.13) shows the fundamental, which corresponds to $\ell = \frac{1}{2}\lambda$ (a single sine hump) and the first three overtones, which correspond to $\ell = \lambda$, $\ell = \frac{3}{2}\lambda$, and $\ell = 2\lambda$ (two, three, and four sine humps). The relative intensities of the fundamental and various overtones determines the timbre of the instrument; a guitar and a violin sound different because, even while playing the same note, the proportions of the overtones differ for the two instruments.[22]

[22] Remind me and we'll get out the old strobe-and-standing-wave demonstration. Always a crowd pleaser.

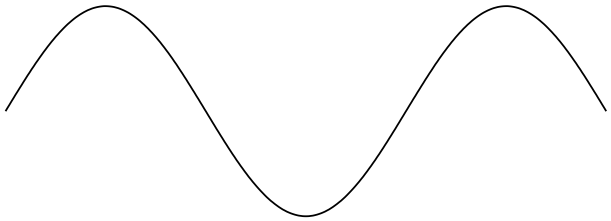


Figure 9.11: The Case $\ell = \frac{3}{2}\lambda$

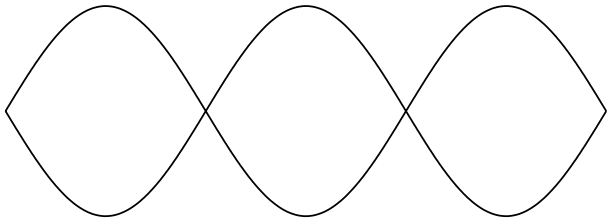


Figure 9.12: The Corresponding Standing Wave

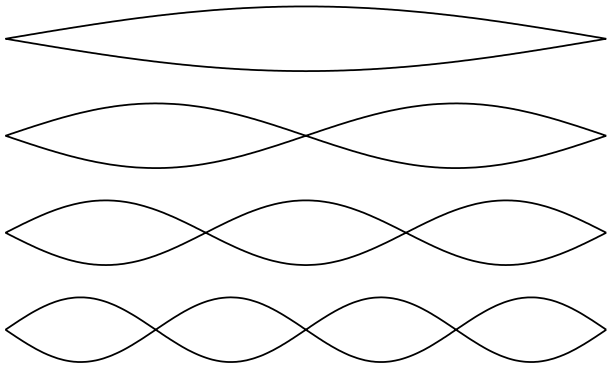


Figure 9.13: The Fundamental and First Three Overtones

9.8.3 Sound Intensity & Decibels

The natural measure for the intensity of sound is power per unit area (in MKS units, $\mathrm{W/m^2}$). The reasoning behind this is as follows: as a first attempt, we might try taking the intensity of a sound to be the energy emitted by a source. But it makes a difference over how much time this energy is emitted: the same quantity of energy emitted over a longer time would not produce as loud a sound. We are thus led to look instead at the *rate* at which the source is emitting sound energy, and power output is the rate of energy output. It also makes a difference over how great an area the sound is distributed: if the same power output is spread out over a greater area, the sound will also not be as loud. Thus we arrive at power per unit area for measuring sound intensity.[23]

Sound sometimes emanates more or less from a point and propagates from that point in all directions, each wave crest traveling outward from the source like a sphere of expanding radius.[24] Since the spherical surface area over which the sound is spread is $4\pi r^2$, the intensity of the sound wave will drop off as $1/r^2$: double your distance from the source, and the sound is only one quarter as loud. Since the area of any shape goes as the square of a linear dimension, there will be a $1/r^2$ drop-off even when, as is commonly the case, the sound intensity is not distributed evenly in all directions. Your stereo speakers, for example, typically emit most of their sound in a forward hemisphere, but the intensity of that sound still drops off as the inverse square of your distance from the speakers.

Your ear is an incredibly sensitive instrument: a person with normal hearing can begin to detect sounds as faint as $10^{-12}\,\mathrm{W/m^2}$. Your ear is also able to sense an incredible range of sound intensity, from the threshold of hearing at $10^{-12}\,\mathrm{W/m^2}$ up to nearly a full $\mathrm{W/m^2}$. To compress this huge range to something more suitable for everyday use, a decibel scale has been defined: if I is the intensity of a sound in $\mathrm{W/m^2}$, the corresponding number of decibels is

$$\text{number of decibels (dB)} = 10\log_{10}\left(\frac{I}{I_0}\right)$$

where $I_0 = 10^{-12}\,\mathrm{W/m^2}$ is taken as the typical threshold of hearing. You can get a feel for the decibel scale from the values in table (9.1).[25]

[23] This is not, however, to say that this physical measure of sound intensity necessarily correlates linearly with the humanly perceived loudness of a sound; the latter involves a lot of physiology as well as physics.

[24] These sorts of spherical harmonic waves are described mathematically by spherical Bessel functions. Just in case you were curious.

[24] Many of these numbers were derived from (*gasp!*) `http://en.wikipedia.org/wiki/Decibel`.

[25] Remind me and we'll get out the decibel meter and see who can scream the loudest. This is always popular with the teachers in neighboring rooms.

dB	Sound
0	Threshold of hearing
1	Republican response to favorable leaks of classified information
2	Mosquito 3 m away
10	Heavy breathing 3 m away
20	Heavy breathing 1 m away
30	Gaffney's fading, distant babbling just before you doze off
40	Residential neighborhood at night
60	Typical close conversation
70	Busy traffic 5 m away
85	Permanent damage with prolonged exposure
90	Lawn mower or heavy truck 1 m away
100	Typical discotech
110	Chainsaw 1 m away
130	Painfully loud
140	Gunshot 1 m away
160	Loud enough to break glass
170	Republican response to unfavorable leaks of classified information

Table 9.1: Typical Decibel Values

9.8.4 Doppler Shift

Any of you who have tried to cross a busy street know that the pitch of a car horn, compared to its pitch when the car is at rest, sounds higher as the car comes toward you and lower after it has passed and is moving away from you.[26] This effect, called the Doppler shift for sound,[27] actually depends on three velocities: the velocity the source of the sound, the velocity of the listener, and the speed at which sound travels through the air. Sound waves are carried by the air, and their speed through it (about 360 m/s, depending on the temperature) does not depend on the motion of either the source of the sound or the listener. But if the source of a sound is moving toward the listener, that source is in effect chasing after its own sound waves, so that successive wave crests get "scrunched up" and the pitch of the sound is increased. Similarly, if the listener is moving toward the source, there is another "scrunching" at the listener's end. Conversely, if the source or the listener is moving away, there is an effective drawing out and lengthening of the wavelength and a corresponding decrease in pitch.

[26] And this effect is independent of whether the driver gives you the finger.

[27] After, oddly enough, some guy named Doppler — the same guy for whom the Doppler shift for light is named. Freaky.

The derivation of the Doppler shift for sound is a little subtle:[28] because the actual stuff of sound waves — the pressure oscillation — exists in and is carried by the air, we want to work in terms of the wavelength of the sound wave in the air and take the motions of the source and the listener into account relative to this. Consider first the emission of the sound by a source S moving toward a listener L. If the source were at rest, the distance between successive wave crests would be cT_s, where c is the speed of sound in air and T_s is the period of the sound wave according to the source. But if the source is chasing after the wave it is emitting, by the time the next wave crest is emitted, the source has moved forward a distance v_sT_s, where v_s is the source's velocity. As a result, the wavelength of the sound in air is reduced to

$$cT_s - v_sT_s = (c - v_s)T_s$$

A similar effect happens at the listener's end: the wavelength that the sound has in the air will be the sum of the "scrunched" wavelength cT_ℓ perceived by the listener and the distance $v_\ell T_\ell$ that the listener has moved "into" the wave between successive wave crests, where T_ℓ is the period of the sound wave according to the listener and v_ℓ is the listener's velocity. The listener's expression for the wavelength of the sound in air is thus

$$T_\ell + v_\ell T_\ell = (c + v_\ell)T_\ell$$

Equating the listener's and source's expressions for the wavelength that the sound has in the air, we have

$$(c + v_\ell)T_\ell = (c - v_s)T_s$$

Inverting both sides and recalling that the frequency f is the reciprocal of the period ($f = 1/T$), we arrive at

$$\frac{f_\ell}{c + v_\ell} = \frac{f_s}{c - v_s} \tag{9.27}$$

Eq. (9.27) is the Doppler shift for sound. When you apply it, be mindful that built into it is the sign convention that v_s and v_ℓ are positive for "approaching" and negative for "receding." If, for example, you run after a teacher and shoot him in the back as he flees from you, the source (you — or, more precisely, the gun) is moving toward the listener (the teacher), so v_s is positive, but the teacher is fleeing from you, so v_ℓ is negative.

If you compare the Doppler shift for sound (eq. (9.27)) to that for light (eq. (10.96)), you see that whereas the Doppler shift for sound depends on the velocities of the source and listener separately (v_s and v_ℓ), the Doppler

[28] *Subtle* is of course just a euphemism for *hard to follow the way we have explained it.*

shift for light depends only on the relative velocity of the source and viewer ($\beta = v/c$). This is because there is a medium that carries sound (the air), with respect to which it is possible to determine absolute velocities for the source and listener. But this is not possible for light: there is no medium carrying light; the æther hypothesized in the late nineteenth century does not exist. Since a fundamental principle of relativity is that there is no preferred frame of reference with respect to which absolute velocities can be determined, and since the source and the listener are the only reference frames involved in the Doppler shift for light, that shift can depend only on the relative velocity of the source and listener.

9.9 Problems

1. Suppose you were attached to a giant spring and oscillating harmonically up and down a long hallway while standing on a skateboard. It's not like stranger things haven't been known to happen. Anyway, where in the oscillation will you feel the greatest effect? That is, at what point will your guts be most churned up?

2. Someone calculates the work done to stretch a relaxed spring a distance ℓ as follows: since

$$F_{spring} = -kx = -k\ell$$

the work done stretching the spring is

$$W = Fd\cos\theta = (k\ell)(\ell)\cos 0 = k\ell^2$$

Explain why this reasoning is boneheaded.

3. You find yourself alone in a room with some stupid object that lies on a frictionless tabletop and oscillates harmonically at the end of a spring according to

$$x = A\cos(\omega t + \varphi)$$

While in the room you have a stopwatch, a meter stick, a graduated cylinder, a six-pack, two slices of leftover pepperoni pizza, a dead camel, and a strange feeling that you are being watched.

   (a) They aren't going to let you out until you determine the values of the parameters A, ω, and φ. Describe how you could do so.

   (b) Determine the values of

$$\frac{|x|}{x_{max}}\,, \qquad \frac{|v|}{v_{max}}\,, \qquad \frac{K}{E}\,, \qquad \text{and} \qquad \frac{U}{E}$$

   where x_{max} and v_{max} are the maximal absolute values of x and v, respectively, when

      i. $|x| = \frac{1}{2}x_{max}$.

      ii. $|v| = \frac{1}{2}v_{max}$.

      iii. $K = \frac{1}{2}E$.

      iv. $U = \frac{1}{2}E$.

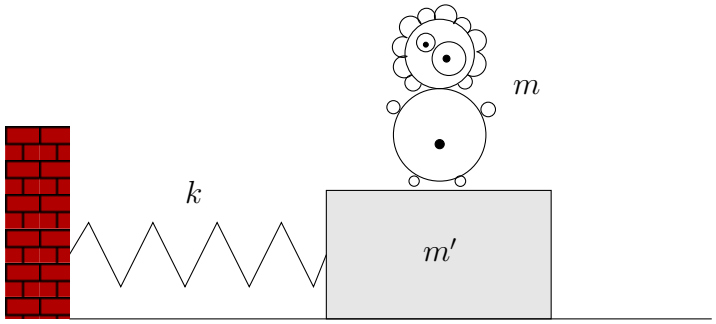


Figure 9.14: Problem 4

4. You devise an improved method for testing the effectiveness of prototype baby powders: the butts of test babies (mass m) are powdered and then placed on a block of rough, splintery wood (mass m') that sits on a frictionless, level surface and is connected to an anchored spring of spring constant k, as shown in fig. (9.14). The amplitude of the spring oscillation is slowly increased until the baby starts to slip on top of the block. It is found that the baby with the smoothest butt slips, falls off, and becomes fouled in the spring when the amplitude of the oscillation has reached the value ℓ. Determine the coefficient of bogus friction between the baby's butt and the block.

5. You (mass m) are chosen as the "test pilot" for a new bungee jump. When you suspend yourself at rest from the bungee, you find that it stretches a distance ℓ under your body weight. The jump point is a height 6ℓ above the ground and the relaxed (unstretched length) of the bungee is 3ℓ. Should you make the test jump?

   Okay, so you know the routine around here well enough to say no without even doing any calculation. But you still need to prove your assertion quantitatively. And when you're doing this, note that *at the jump point* the bungee is totally slack and therefore effectively not even a part of the problem; the "springiness" of the bungee doesn't become relevant until you reach the 3ℓ-mark and it begins to stretch.

6. You've probably often wondered what would happen if you fell into a hole that went "all the way to China."[29] Suppose we idealize the problem by imagining that a tunnel, narrow enough that its volume is negligible, is bored through a sphere of uniform mass density ρ. Show that the gravitational pull exerted on you is $\frac{4}{3}\pi G m \rho r$, where r is your distance from the center of the sphere, and that you therefore oscillate harmonically.

[29] Do people in China speak of holes that go all the way to the United States? Or is this all now considered politically incorrect, and we're supposed to speak of holes that go "all the way to the lands of people whose cultural perspectives are equally valid"?

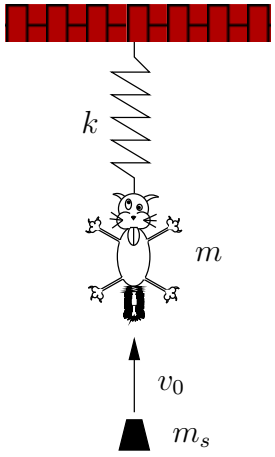


Figure 9.15: Problem 7

7. A standard test cat (mass m) is suspended at rest from a vertical spring of spring constant k. A rubber stopper of mass m_s, fired straight upward, strikes the cat at speed v_0 and lodges in it, as shown in fig. (9.15).

   (a) Determine the amplitude of the resulting oscillation.

   (b) Make physical sense of the dependence of your result on v_0, k, m_s, m, and, in particular, the quantity $m_s g/k$.

8. (The Return of the Weird Spring.) Recall that the weird spring in #35 on p.288 exerted a force

$$F = kx - \beta x^3$$

   where k and β were positive constants and x measured the spring's distention or compression. In #35 on p.288 you found that the potential energy associated with this spring force was

$$U = -\tfrac{1}{2}kx^2 + \tfrac{1}{4}\beta x^4$$

   (a) Determine the locations of the stable equilibrium points of this potential.

   (b) A particle of mass m at one of these stable equilibrium points is given a slight nudge to set it in motion. Determine the angular frequency of the resulting oscillation.

   (c) Suppose that, instead of being anchored at one end, the spring connects two masses, m_1 and m_2, so that they exert this weird spring force on each other (but do not experience any other forces). Determine the angular frequency of small oscillations.

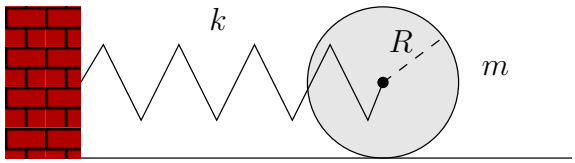


Figure 9.16: Problem 9

9. The free end of a spring of spring constant k is connected to the axle of a uniform solid cylinder of mass m and radius R, as shown in fig. (9.16). The cylinder is released from rest when the spring is compressed and rolls without slipping along a level surface.

   (a) Assuming that the resulting oscillations are harmonic, show that the angular frequency of small oscillations is

   $$\omega = \sqrt{\frac{2k}{3m}}$$

   (b) Is it necessary that the oscillations be small, or will they be harmonic for any amplitude?

   (c) Show, by considering kinetic and potential energies, that the oscillations are indeed harmonic. See the footnote if you need a hint.[30]

   (d) Show, by considering forces and torques, that the oscillations are indeed harmonic.

[30] How do the expressions for the kinetic and potential energies of the rolling cylinder compare to those you would have if the cylinder were instead sliding without friction? You should find that your expression for energy of the rolling cylinder can be made identical to that of the sliding cylinder if you regard the rolling cylinder as having an "effective" mass that is a multiple of its true mass m.

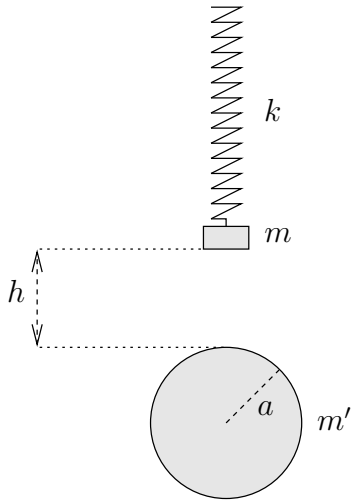


Figure 9.17: Problem 10

10. A boring mass m is attached to the end of an equally boring spring of spring constant k and suspended over a uniform spherical asteroid of mass m' and radius a, as shown in fig. (9.17). The asteroid is also boring. When the mass m is hanging at rest, it is a distance h above the surface of the asteroid. How the "upper" end of the spring is supposed to be anchored in empty space is anybody's guess, but, while we're at it, we'll suppose that the asteroid is somehow magically fixed in place too. Also note that, as implied by the scale of fig. (9.17), h is comparable in size to the radius a of the asteroid. Hint, hint. Wink, wink.

    (a) Determine the period of small oscillations of the mass m.

    (b) What limits of the parameters m, m', k, h, or a can you use to check your result? Show that you get what you expect in these limits.

    (c) If the asteroid is instead free to move under the influence of the gravitational force between it and the mass m, should m be replaced by the reduced mass

$$\mu = \frac{mm'}{m + m'}$$

    in your result for the period?

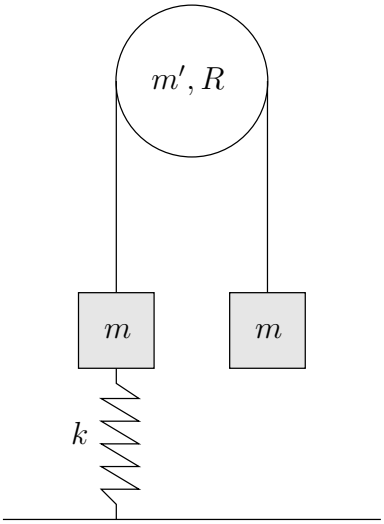


Figure 9.18: Problem 11

11. Fig. (9.18) shows two blocks, both of mass m, slung over a frictionless pulley that is a uniform disk of mass m' and radius R. One of the masses is connected to an ideal spring of spring constant k. Initially, everything is at rest and the spring is relaxed. You then pull the mass on the left side down, compressing the spring a distance ℓ, and release it from rest. Try to contain your excitement.

    (a) What is the period of the resulting oscillation? Note that very little calculation is required. See the footnote if you need a hint.[31]

    (b) Is your result for the previous part valid for oscillations of any size or only for small oscillations? (Do not worry about the blocks hitting the ground or the pulley.)

[31] You will want to think about how much mass is, effectively, in motion during the oscillation.

12. The energy E of a mass m moving in the gravitational field of a humongous (and therefore effectively stationary) mass M at the origin is

$$E = \tfrac{1}{2}mv^2 - \frac{GMm}{r}$$

Recall now (eq. (3.41) on p.151) that in radial coordinates the expression for the velocity v is

$$v^2 = \dot{r}^2 + r^2\dot{\theta}^2$$

where a dot denotes a time derivative. We can rewrite this v in terms of the angular momentum $L = I\omega = mr^2\dot{\theta}$ by substituting L/mr^2 for $\dot{\theta}$:

$$v^2 = \dot{r}^2 + r^2\left(\frac{L}{mr^2}\right)^2 = \dot{r}^2 + \frac{L^2}{m^2r^2}$$

The energy E thus becomes

$$E = \tfrac{1}{2}m\dot{r}^2 + \frac{L^2}{2mr^2} - \frac{GMm}{r}$$

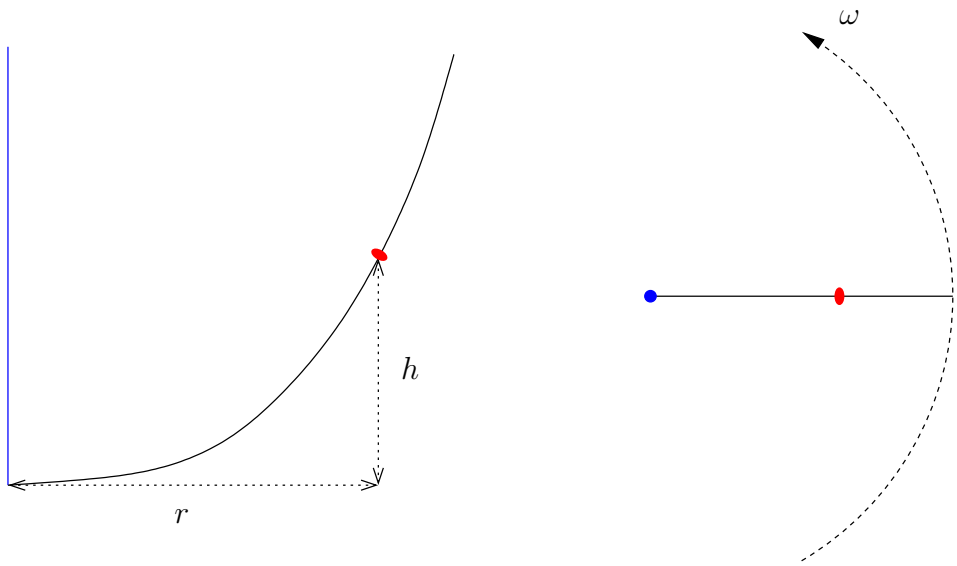
Since a central force does not exert any torque, the angular momentum L is constant, and we now have an expression for E that is a function of the single variable r and that we can regard as the sum of a radial kinetic energy $\frac{1}{2}m\dot{r}^2$ and an effective potential

$$U_{\mathit{eff}} = \frac{L^2}{2mr^2} - \frac{GMm}{r}$$

The term $L^2/2mr^2$ is called the centrifugal potential.

(a) Show that minimizing the effective potential with respect to the radius r of orbit gives the usual expression for the radius of a circular orbit. (That is, gerrymander the result that this minimization gives you for the radius into a form that is easily seen from setting the centripetal and gravitational forces equal.)

(b) From the effective potential, calculate the period of small oscillations about the stable equilibrium radius, that is, the period of small oscillations in the circular orbit. Show that the period of these small oscillations is the same as the period of the circular orbit.



View from the side View from above

Figure 9.19: Problem 13

13. Fig. (9.19) shows a piece of wire (the solid black curve) bent upward into a curve $h(r)$, where r is the radial distance, and h the height, from the bottom end of the wire. An essentially point-like bead of mass m (shown in red) is mounted on the wire and free to slide frictionlessly along it. The wire is then rotated around a vertical axis (shown in blue) at a constant angular velocity ω.

    (a) Determine the radius r at which the bead will be in equilibrium if the curve of the wire is the straight line $h = cr$, where c is a positive constant.

    (b) Determine the radius r at which the bead will be in equilibrium if the curve of the wire is $h = cr^3$, where c is a positive constant (though of course different from the c of the preceding part).

    (c) Determine the radius r at which the bead will be in equilibrium if the curve of the wire is $h = cr^2$, where c is a positive constant. You may find your result surprising at first, but on reflection it should make sense. If not, keep reflecting.

    (d) If we were, when setting up the relation $F_{radially\ inward} = m\omega^2 r$ for the circular motion of the bead, to move the $m\omega^2 r$ from the ma side of $F = ma$ to the F side and regard it as the contribution of a fictitious radially inward force $-m\omega^2 r$ — that is, to a radially *outward* centrifugal force

$$F_{centrifugal} = +m\omega^2 r$$

— , then we could, because ω is constant, ascribe this centrifugal force to a centrifugal potential [32]

$$\begin{aligned} U_{centrifugal} &= -\int_0^r F_{centrifugal}\, dr \\ &= -\int_0^r m\omega^2 r\, dr \\ &= -m\omega^2 \int_0^r r\, dr \\ &= -\tfrac{1}{2} m\omega^2 r^2 \end{aligned}$$

Using this centrifugal potential together with the gravitational potential energy mgh, determine for each of the three preceding parts whether the equilibrium point is stable or unstable. If it is stable, determine the angular frequency of small oscillations about it. See the footnote if you need a hint.[33]

(e) Make physical sense of the signs of your results for the effective spring constants in the preceding part. See the footnote if you need a hint.[34]

[32] Note that this result for the centrifugal potential differs from that of problem #12 because in this problem it is ω, not the angular momentum L, that is constant: as the bead moves along the wire, the wire can exert on the bead a horizontal force that will provide whatever torque is needed to keep the angular velocity ω constant.

[33] While m, g, and ω are constants, h is a function of r, which is varying. You will therefore need to use $h = h(r)$ to rewrite the gravitational and hence the total potential energy in terms of r before determining the effective spring constant.

[34] You will want to think in terms of the behaviors of the required centripetal force and of the radial component of the normal force as functions of r.

14. Tarzan (mass m), who is romantic but not too swift, swings from rest on a long vine of length ℓ down to Jane (also mass m). Jane, who is also not too swift, awaits Tarzan with open arms. Jane is at the bottom of the swing, and the vine initially makes an angle $\theta_0 = 60°$ with the vertical (which we will, with a reckless disregard for truth, treat as a small enough angle that the swing is harmonic).

    (a) If possible, determine how long it takes Tarzan to reach Jane.

    (b) If possible, determine how long it takes Tarzan to travel the first $\frac{1}{2}\theta_0$ toward Jane.

    (c) If possible, determine how long it takes Tarzan to travel the second $\frac{1}{2}\theta_0$ toward Jane.

    (d) Tarzan and Jane stick together as a result of the collision. As this combined mass swings upward after the collision, how long does it take them to come to rest?

    (e) If the vine is 10 m long, roughly how fast is Tarzan moving when he reaches Jane? Will the experience be conducive to romance?

15. In working out #14e, someone argues that since the pendulum motion is approximately harmonic with angular frequency $\omega = \sqrt{g/\ell}$, the angle θ of orientation as a function of the time t is given by

$$\theta = \theta_0 \cos \omega t = \theta_0 \cos\left(\sqrt{\frac{g}{\ell}}\, t\right) \tag{9.28}$$

    so that the speed at which tarzan travels along the arc is

$$v = \frac{ds}{dt} = \frac{r\, d\theta}{dt} = \ell \frac{d\theta}{dt} = \ell \frac{d}{dt}\left(\theta_0 \cos\left(\sqrt{\frac{g}{\ell}}\, t\right)\right) = -\theta_0 \sqrt{g\ell}\, \sin\left(\sqrt{\frac{g}{\ell}}\, t\right)$$

    This satisfies $\theta = \theta_0$ and $v = 0$ when Tarzan begins his descent at $t = 0$. When he reaches Jane, $\theta = 0$, which, in eq. (9.28), yields

$$t = \frac{\pi}{2}\sqrt{\frac{\ell}{g}}$$

    and thus

$$v = -\theta_0 \sqrt{g\ell} \sin\frac{\pi}{2} = -\frac{\pi}{3}\sqrt{g\ell}$$

    What's up with that?

16. In working out #14a, someone reasons as follows: Since Tarzan starts from rest, his speed when he reaches the bottom of the swing is, by conservation of energy, given by

$$\tfrac{1}{2}mv^2 = mgh = mg\ell(1-\cos\theta_0)$$

which yields

$$v = \sqrt{2g\ell(1-\cos\theta_0)}$$

And since the distance covered by Tarzan is the arc length $s = \theta_0\ell$, the relation $x - x_0 = \frac{1}{2}(v + v_0)t$ becomes

$$\theta_0\ell = \tfrac{1}{2}\left(\sqrt{2g\ell(1-\cos\theta_0)} + 0\right)t$$

which yields for the time t that it takes Tarzan to reach Jane

$$t = \theta_0\sqrt{\frac{2\ell}{g(1-\cos\theta_0)}}$$

Explain why this reasoning is boneheaded.

17. Someone reasons that since the angular frequency of a simple pendulum is

$$\omega = \sqrt{\frac{g}{\ell}}$$

and the angular frequency of harmonic oscillations more generally is

$$\omega = \sqrt{\frac{k_{\mathit{eff}}}{m}}$$

the effective spring constant for a simple pendulum must be

$$k_{\mathit{eff}} = \frac{mg}{\ell}$$

If the motion of the pendulum is expressed in angular terms, conservation of energy then gives

$$\tfrac{1}{2}I\omega^2 = \tfrac{1}{2}k_{\mathit{eff}}\theta_{max}^2$$
$$\tfrac{1}{2}(m\ell^2)\left(\sqrt{\frac{g}{\ell}}\right)^2 = \tfrac{1}{2}\left(\frac{mg}{\ell}\right)\theta_{max}^2$$

which yields

$$\theta_{max} = \ell$$

Explain why this reasoning is boneheaded.

18. Suppose you are obsessed with grandfather clocks (the timing mechanisms of which, for the benefit of those of you with limited familiarity with grandfather clocks, are pendula).

    (a) You take a grandfather clock that keeps correct time into an elevator that starts falling at $\frac{1}{3}g$. Will the clock gain or lose time? That is, will it run fast or slow?

    (b) How many seconds per hour will the clock gain or lose?

    (c) After you get off the elevator, you take the clock into orbit on the Space Shuttle. Will it gain or lose time? How many seconds per hour?

    (d) Assuming that the Space Shuttle makes it back, you plan to bring the clock to the North Pole, calibrate it, and then take it to the Equator. At the Equator, will it gain or lose time? How many seconds per hour? With a persistent reckless disregard for the truth, we will take the Earth to be a sphere of radius $\approx 6.38 \times 10^6$ m.

19. Walking can in some approximation be regarded as pendulum motion, with the forward stride of each leg likened to the forward swing, solely under the influence of gravity, of a uniform rod pivoted stiffly about the hip joint. This technique has been used to estimate the walking speed of various dinosaurs. Use it to estimate human walking speed.

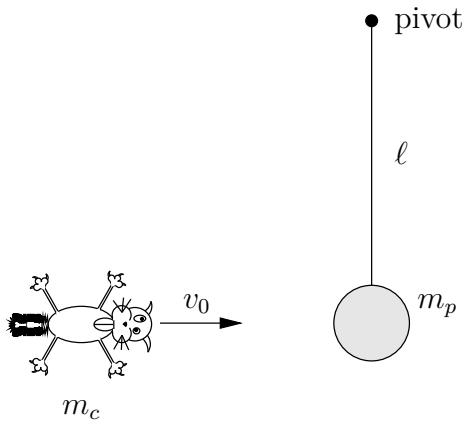


Figure 9.20: Problem 20

20. At a top-secret military research center, a cat of mass m_c is fired into a pendulum that consists of a bob of mass m_p at the end of a massless rod of length ℓ, as shown in fig. (9.20). The pendulum is initially at rest, and the cat is moving with a horizontal velocity v_0 when it strikes and sticks to the pendulum.

    (a) Show that when the pendulum bob and the cat are, for simplicity, both treated as a point masses, the amplitude of the resulting swings of the pendulum, as an angle θ_{max} with the vertical, is given by

$$\cos\theta_{max} = 1 - \frac{v_0^2}{2g\ell}\left(\frac{m_c}{m_c + m_p}\right)^2$$

    (b) Suppose that the pendulum now consists, not of a point mass m_p at the end of a massless rod of length ℓ, but rather of a uniform thin rod of mass m_p and length ℓ. Show that the amplitude of the resulting swings of the pendulum, as an angle θ_{max} with the vertical, is now given by

$$\cos\theta_{max} = 1 - \frac{v_0^2}{2g\ell}\left(\frac{m_c^2}{\left(m_c + \frac{1}{3}m_p\right)\left(m_c + \frac{1}{2}m_p\right)}\right)$$

    For simplicity, continue to treat the cat as a point mass. See the footnote if you need a hint.[35]

[35] Consider carefully what quantity is conserved in the collision. You could get away with not thinking much about this in the preceding part, but not in this part.

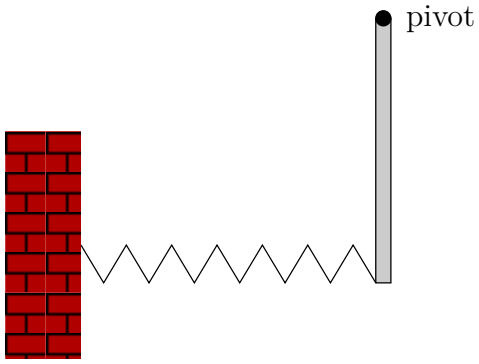


Figure 9.21: Problem 21

21. The free end of a spring of spring constant k is connected at a right angle to one end of a uniform thin rod of length ℓ and mass m, as shown in fig. (9.21). The rod pivots freely about its other end like a pendulum.

    (a) Show that the angular frequency of small oscillations is

    $$\omega = \sqrt{3\left(\frac{k}{m} + \frac{g}{2\ell}\right)}$$

    It may help to remember that $1 - \cos\theta \approx \frac{1}{2}\theta^2$ for small θ, or that $\sqrt{1-\epsilon} \approx 1 - \frac{1}{2}\epsilon$ for small ϵ. See the footnote if you need a hint.[36]

    (b) Make physical sense of this result in the limits

    i. m is small.
    ii. m is large.
    iii. k is small.
    iv. k is large.
    v. ℓ is small.
    vi. ℓ is large.

    (c) If you did #21a by forces and torques, now do it by looking at kinetic and potential energies.

    (d) If you did #21a by energy, now do it by forces and torques.

[36] Although the motion is along a circular arc, to first order for small oscillations the arc can be treated as a straight line; the deviation from a straight line will be second-order small. You will also want to think in terms of the expression for the energy and of effective mass, as you did in #9 and #11.

22. For some reason, you feel it necessary to have in your dorm room a stereo that can crank out 100 watts of sound.

    (a) Under the rather bogus but simplifying assumption that the speakers emit sound evenly over the forward hemisphere, what is the intensity of the sound in $\mathrm{W/m^2}$ at 3.0 m from the speakers? (Assuming as well that your dorm room is big enough that it's possible to be 3 m from the speakers.)

    (b) With how many decibels are you therefore getting blasted when you are 3.0 m from the speakers?

    (c) If, in addition to the usual drop in intensity with distance, the intensity of the sound is reduced by a factor of 10 by the intervening walls, how many decibels are heard by the housemaster, who is 50 m away?

23. On a visit to Graceland with the Bushes, Japanese Prime Minister Junichiro Koizumi starts belting out Elvis Presley tunes in the Jungle Room.[37] If you are catching a 90 dB earful 4.0 m away from Koizumi, how many watts is Koizumi cranking out? For simplicity, make the bogus assumption that Koizumi emits sound evenly in all directions.

24. While running across the tops of the dining-hall tables at $\frac{1}{6}c$ to get your seventh plate of buffalo wings, you eruct vigorously at a dominant frequency of 70 Hz. Using $c = 360$ m/s for the speed of sound,

    (a) Determine the frequency heard by people sitting in front of you.

    (b) Determine the frequency heard by people sitting in behind you.

25. How fast would a school bus have to bear down on a small child in order that the child's 1000 Hz shriek be out of the driver's range of hearing? The range of human hearing varies considerably among individuals, and sensitivity to high frequencies diminishes with age, but a person with good hearing can typically hear sounds from 20 to 20,000 Hz. You may again take the speed of sound to be 360 m/s. Also assume that the child is paralyzed with fear.

[37] We wish we were making this up, but, believe it or not, it really happened.

9.10 Sketchy Answers

(4) $\dfrac{k\ell}{(m+m')g}$.

(7a) $A = \dfrac{m_s g}{k}\sqrt{1 + \dfrac{kv_0^2}{(m+m_s)g^2}}$.

(8a) $x = \pm\sqrt{\dfrac{k}{\beta}}$.

(8b) $\sqrt{\dfrac{2k}{m}}$.

(10a) $2\pi\left(\dfrac{k}{m} - \dfrac{2Gm'}{(h+a)^3}\right)^{-\frac{1}{2}}$.

(11a) $2\pi\sqrt{\dfrac{2m + \frac{1}{2}m'}{k}}$.

(13a) $\dfrac{cg}{\omega^2}$.

(13b) $\dfrac{\omega^2}{3cg}$.

(13d) For the only stable equilibrium, $\omega_{osc} = \omega$.

(18b) 661 sec.

(18d) 6.20 sec.

(22a) 1.77 W/m^2.

(22b) 122 dB.

(22c) 88 dB.

(23) 0.20 W.

(24a) 84 Hz.

(24b) 60 Hz.

(25) 6840 m/s — which, inasmuch as this is 19 times the speed of sound, isn't possible. You'll just have to plug your ears.

Part III

Beyond Basic Mechanics

Chapter 10

Relativity

Da könnt mir halt der liebe Gott leid tun, die Theorie stimmt doch.

Albert Einstein [1]

There was a young lady named Bright
Who traveled much faster than light.
She started one day
In the relative way
And returned on the previous night.

A.H.R. Buller

10.1 Reference Frames

Reference frame is simply a technical term for the perspective of an observer. For our present purpose, these reference frames differ from each other by their relative motion: the same physical events may be watched by observers in several different reference frames; because of their relative motion, these observers will agree on the values of some physical quantities and disagree on the values of others. *Relativity*, in the sense in which we will be using it, refers to these differences in perspective and in the values that the observers report for various physical quantities.[2]

[1] This was Einstein's reply when asked how he would react if the 1919 expedition to measure the bending of starlight by the Sun's gravitational pull had contradicted his general theory of relativity. Translation: "Then I would have felt sorry for the dear Lord; the theory is correct."

[2] Actually, the quantities that have physical meaning are those that (like total energy) do not change. Such quantities are said to be conserved or invariant. When Einstein first proposed his theory of relativity, he debated with himself whether to call it *invariance*

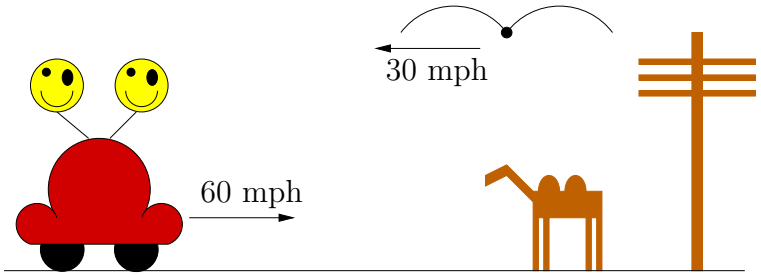


Figure 10.1: Look Out for That Camel!

Suppose, for example, that you and your friend are driving north at 60 mph as a bird flies south at 30 mph and a two-humped camel watches everything as it stands at rest by the side of the road, as shown in fig. (10.1). Or rather, as we have attempted to show in fig. (10.1), with what looks like Siamese twins driving the Oscar Meyer Wiener Mobile, a very cubist representation of a camel, and a large pterodactyl. Anyway, we have four observers in three different reference frames: the bird, the camel, and you and your friend (who, along with the car and everything moving with the car, share the same reference frame). According to the kind of common-sense relativity we've been using up to now — the kind on which Newtonian physics is based[3] — , all four observers would agree on the values of some quantities, such as the time it takes you to go between two road signs or the distance between those signs, while they would disagree on others, such as the velocity of the phone poles: to the camel, who is at rest by the side of the road, the phone poles of course seem to be at rest. But to you, who are driving north at 60 mph, the phone poles appear to be moving south at 60 mph, and to the bird, who is flying south at 30 mph, they appear to be moving north at 30 mph. Similarly to you the bird seems to be flying south at a "combined" speed of $60 + 30 = 90$ mph, while the bird, to itself, seems to be at rest.[4]

theory (to emphasize those quantities that are the same for different observers) or *relativity theory* (to emphasize those quantities that differ for different observers). He decided on the latter but later regretted this choice because of the rampant popular abuse of the term — everyone going around making the most stupifyingly incredible arguments based on the premise that "everything is relative," as though this were an indisputable and ironically absolute law that can be applied at all times in every possible context.

[3]Oddly enough, this sort of relativity is known, not as Newtonian relativity, but as Galilean relativity.

[4]This is an important point of which you should make note: an object is always at rest from its own perspective. In some ways, this is the physics version of the sayings "you take yourself with you wherever you go" and "wherever you go, there you are." Since the reference frame of an object is always moving with the object, an object is always at rest in its own reference frame.

When we work with physical quantities relativistically, it will therefore be very important to specify not only their values, but also according to which observers (that is, in which reference frames) they have those values.

10.2 Einstein Disses Newton

Reference frames fall into two basic classes: inertial and noninertial. Inertial frames are those that do not have any acceleration; noninertial frames are those that do. If we treat the Earth as though it were at rest (which is essentially what we have done up to now when working out projectile and other kinds of motion), then it constitutes an inertial frame, as do all objects at rest on it and, their acceleration being zero, all objects moving at constant velocity relative to it. If, however, we take into account the spin of the Earth about its own axis,[5] then the Earth and objects at rest on it are, by virtue of their circular motion, experiencing the usual centripetal acceleration and therefore constitute noninertial frames. And though it would be somewhat more complicated to work out,[6] objects moving at constant velocity relative to the Earth would have a similar acceleration and likewise constitute noninertial frames. Whether for the purposes of a particular application like projectile motion we can ignore the Earth's spin about its own axis is a question of approximation: strictly speaking, the Earth constitutes a noninertial frame, but its acceleration is rather small and for many practical applications can to a good approximation be ignored.

In Newtonian physics the distinction between inertial and noninertial frames is critical: in inertial frames, $F = ma$ is obeyed; in noninertial frames, because of the apparent contribution to the ma side of $F = ma$ from the frame's own acceleration, it is not. You are familiar with this from common experience: when you go around a circular turn in a car, you and the car are experiencing the usual centripetal acceleration v^2/r, so your frame is noninertial. This noninertiality manifests itself as the *centrifugal* force you subjectively perceive: you feel yourself being thrown to the outside of the turn. In fact, the reality is exactly the reverse: in order for you to stay in the car's frame and make the turn with the car, you need some physical force — usually friction between you and the seat — to pull you in toward the center of the turn and supply the necessary *centripetal* force. In the absence of this required centripetal force, you would simply move at a constant velocity in a straight line tangent to the car's circular path (as was shown in fig. (4.7) on p.206); there isn't any force throwing you to the outside of the turn, and this perceived centrifugal force is therefore termed fictitious.

[5] Or its orbit around the Sun, or the orbit it shares with the Sun and the rest of the solar system about the center of the galaxy, etc.

[6] Actually, we did work this out, in §7.3.

The relativity of motion in Newtonian physics therefore rests on a single postulate: that the laws of physics (like $F = ma$) are the same in all inertial frames. That is, as long as an observer's reference frame does not have any acceleration that would mess up $F = ma$ and other physical laws by introducing fictitious forces, everything seen by that observer should be the same as for an observer at rest. As an illustration, consider taking a sip out of a hot cup of coffee: if you take a sip while riding in a car moving at a constant velocity — no changes in speed or direction — , then nothing will seem (or in fact be) at all different from when you take a sip while at rest on the ground.[7] If, however, the frame of the car becomes noninertial by slowing down, speeding up, or making a turn, weird things happen. In particular, if the car suddenly speeds up while you are trying to take a sip, the coffee in the cup does not share that forward acceleration and therefore seems, from the perspective of objects like you that do share it, to be sloshed backward, with the result that sharp words with the driver ensue.

Now, sound waves are carried by the medium of the air — at, depending on the temperature and other conditions, about 360 m/s. As you would expect, if you approach the source of a sound at 30 m/s, the sound wave, as "seen" by you, approaches at $360 + 30 = 390$ m/s. Similarly, if the wind is blowing at 10 m/s, sound waves are carried with the moving air and share this 10 m/s velocity, so that, depending on their direction of motion, these waves will travel at anywhere from $360-10 = 350$ m/s to $360+10 = 370$ m/s. When people began investigating similar effects in light waves back in the late 19th century, there was speculation that there must be some as yet unobserved medium that carried light waves in the same way that the air carries sound waves. They called this medium the æther and set up experiments to detect the effects of motion through the æther on the speed of light waves. Since the speed of light is so large ($c \approx 3.00 \times 10^8$ m/s), motion of very high velocity was required; the experiments used the orbital velocity of the Earth around the Sun. While this velocity is still much smaller than the speed of light, it is large enough that any changes in the speed of light should have been observable. The first and most famous of these attempts to detect the æther was the Michelson-Morley experiment, in which mirrors and half-mirrors were used to split a light beam and bounce it back and forth along two paths, one parallel to the Earth's orbital motion and the other perpendicular to it. The effect of the Earth's motion through the æther on the light waves following these two paths was expected to be just like the headwind-tailwind and crosswind parts of problem #83 on p.184 and should, when the waves were subsequently recombined, have resulted in a measurable shift in the

[7] You might object that there are still differences due to bumps and vibrations in the ride, but remember that those constitute accelerations. A truly constant velocity would mean a perfectly smooth ride.

interference pattern they produced. In fact, no effect beyond the bounds of experimental error was detected, neither in the original Michelson-Morley experiment nor in the many more accurate such experiments that have since been carried out.[8] As much as it violated common sense, the only possible conclusion seemed to be that the speed of light is the same in all reference frames, regardless of their relative motion.

Along with this experimental evidence, there were even more compelling theoretical reasons for believing that the speed of light was constant regardless of the motions of the source or observer: Over the course of the 19th century, as people investigated the great variety of electric and magnetic effects, relations among these seemingly disparate effects were gradually found. In the 1860s, the great Scottish physicist James Clerk Maxwell realized that in fact all of the great diversity of electromagnetic phenomena were accounted for by just four equations, known today as the Maxwell equations:[9]

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\rho \qquad\qquad \nabla \cdot \mathbf{B} = 0$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \qquad\qquad \nabla \times \mathbf{B} = \mu_0 \mathbf{j} + \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t}$$

If you know enough math, you can combine these equations to generate a differential equation that has a wave solution and thus predict the existence of light waves, including the result $c = 1/\sqrt{\epsilon_0 \mu_0}$ for the speed c of light in terms of two fundamental electromagnetic constants, the electric permittivity ϵ_0 and magnetic permeability μ_0 of the vacuum. Curiously, although the Maxwell relations are ultimately based on experimental observations, there is no reference in them to the frame of any particular observer, and one is therefore led to conclude that electromagnetic phenomena, including the speed of light, are independent of the observer and the same in all reference frames.

Einstein therefore based his theory of relativity on two postulates:[10]

[8] Those interested can find Michelson's and Morley's original paper at `http://www.aip.org/history/gap/PDF/michelson.pdf`.

[9] Actually, the way Maxwell himself expressed the Maxwell equations was very messy; the equations were later re-expressed in their current simplified form by Oliver Heaviside.

[10] In his famous 1905 first paper on relativity in *Annalen der Physik*, "Zur Elektrodynamik bewegter Körper" (On the Electrodynamics of Moving Bodies), the words with which Einstein himself stated the two postulates were

1. Die Gesetze, nach denen sich die Zustände der physikalischen Systeme ändern, sind unabhängig davon, auf welches von zwei relativ zueinander in gleichförmiger Translationsbewegung befindlichen Koordinatensystemen diese Zustandsänderungen bezogen werden.

2. Jeder Lichtstrahl bewegt sich im „ruhenden" Koordinatensystem mit der bestimmten Geschwindigkeit V, unabhängig davon, ob dieser Lichtstrahl von einem

- The laws of physics are the same in all inertial frames.
- The speed of light is the same in all reference frames.

The first of these postulates is shared with Newtonian physics; it is the second that makes all the difference: Newtonian physics and common sense would lead you to believe that if you are driving down the road at 60 mph and turn on your headlights, the light coming out of those headlights would be traveling at the usual $c = 3.0 \times 10^8$ m/s plus 60 mph. Einstein's theory says that in fact the speed of the light from those headlights is the same 3.0×10^8 m/s that it would be if the car were at rest, or indeed moving at any other velocity. And we all know what an awesomely righteous dude Einstein was.

In the sections that follow, we'll work out some of the consequences of this second postulate. Since the postulate itself is contrary to common sense, many of its consequences also turn out to be contrary to common sense.[11] These relativistic effects have, however, all been abundantly verified experi-

---

ruhenden oder bewegten Körper emittiert ist. Hierbei ist

$$\text{Geschwindigkeit} = \frac{\text{Lichtweg}}{\text{Zeitdauer}} ,$$

wobei „Zeitdauer“ im Sinne der Definition des §1 aufzufassen ist.

The first of these is clumsy to translate directly into English, but they may be paraphrased as

1. The laws by which the states of physical systems evolve are independent of which of two coordinate systems in uniform motion relative to each other this evolution is formulated in terms of.

2. Any light ray moves in the stationary coordinate system with the fixed velocity c, independent of whether this light ray is emitted by a stationary or moving body. Hence

$$\text{velocity} = \frac{\text{light path}}{\text{time interval}} ,$$

where “time interval” is in the sense of the definition of §1 [referring to a discussion of time intervals and light rays in the previous section of the paper].

Maybe not stated as elegantly as they might have been with greater consciousness of their subsequent historical importance, but there you have it.

Those interested might check out the discussion of the second postulate in *The Function of the Second Postulate in Special Relativity*, Alon Drory, `https://arxiv.org/abs/1408.3142` and the references therein.

[11] Common sense really isn't all it's cracked up to be.

mentally, and logically the theory is completely self-consistent.[12,13]

Much of the beauty of relativity theory lies in its simplicity: it's all based on just the two simple postulates above. Strictly speaking, the theory based on the above postulates is called special relativity, because the first postulate is restricted to the special case of inertial reference frames. Though it requires a great deal more sophisticated math to work out, it is possible to drop the complications of that restriction and formulate an even simpler theory in which the first postulate applies to all reference frames, whether inertial or noninertial. Because the first postulate is no longer a special case, this theory is known as general relativity. We'll say a little about general relativity when we finish our study of special relativity; it turns out that, in addition to other nice results, a theory of gravity just falls out of general relativity.

10.3 The Lorentz Transform

Like any other velocity, the speed of light is calculated by dividing the distance the light moves by the corresponding time interval. If the speed of light, contrary to naive expectations, is to be the same in all reference frames, then something weird must therefore happen to spatial distances and time intervals.

The relations that will turn out to govern this weirdness and to enable us to relate times and distances in one reference frame to times and distances in another reference frame are known as the Lorentz transform.[14] We will

[12]When people doubt the validity of relativity theory, it is usually for disingenuous egotistical reasons. Some people get upset that the universe doesn't work the way they think it should or that it works in ways they find hard to understand. Presumably these were the same people who, a generation or two before, objected to the notion that the Earth was round, that the Sun is the center of the solar system, that humans and apes evolved from a common ancestor, and that much of what goes on inside our heads is not only unconscious but beyond our conscious control. Humans are separated from other animals principally by their capacity for irrationality. As a species, we have a deeply egotistical need to believe that we and our place in the universe are somehow special.

[13]Sometimes you will read about so-called paradoxes in relativity theory. All these "paradoxes" show is that if you think nonrelativistically, you don't get the relativistic answer — not, on the face of it, a terribly surprising result. In fact, such paradoxes are often presented ironically in textbooks as exercises to improve one's understanding of relativity theory. We will deal with a couple of these in the problems at the end of the chapter.

[14]This of course begs the question of why it's called the *Lorentz transform* and not the *Einstein transform*. H.A. Lorentz actually began proposing his Lorentz transform in 1895, ten years before Einstein formulated special relativity, but he lacked Einstein's profound insight into the physical basis for relativity; his work was more of an *ad hoc* attempt to make Newtonian physics consistent with the notion of æther in spite of the null result of the Michelson-Morley experiment and also to account for some observations, made by Walter Kaufmann from 1901-1906, of the electron's apparent mass as a function of its velocity. Einstein, in his special theory of relativity, was the first to present a comprehensive understanding of the Lorentz transform as a fundamental property of spacetime. (Though

ultimately give four different derivations of the Lorentz transform: After we introduce relativistic effects in the next section by working out the consequences of the constantness of the speed of light for a light pulse bouncing off of a mirror, in §10.3.2 we will derive the Lorentz transform based on this and other concrete cases. Then in §10.3.3 we will give a bit more abstract but cleaner and more elegant derivation. Our third derivation, in §10.3.4, is more in keeping with the modern understanding of relativity theory. Finally, in §10.3.7 we will derive the Lorentz transform in a more sophisticated way from spacetime symmetry. Overkill? Maybe.[15] But you can never have too much of a good thing. And the Lorentz transform is after all at the heart of special relativity.

10.3.1 Special Case: A Mirror & Light Pulse

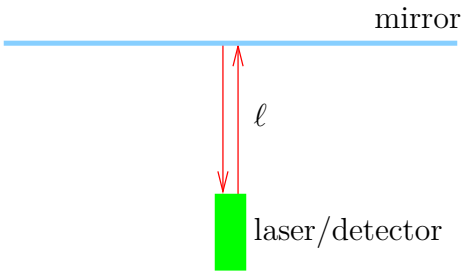


Figure 10.2: Light Pulse in the Lab Frame

To quantitatively investigate the weirdness that happens to spatial distances and time intervals, we begin by considering a simple special case: the firing of a light pulse from a laser at a right angle to a mirror and the subsequent reception of that pulse upon its return to the laser's location. As shown in a view from above in fig. (10.2), in the frame in which the laser and mirror are at rest, the light pulse simply doubles back on itself. (In the

in fairness we should note that there is still today some dispute about who among the early pioneers of relativity theory — Einstein, H.A. Lorentz, and Henri Poincaré — had priority for which aspects of the theory. There is also, to a lesser extent, some squabbling over how credit for the formulation of the field equations of general relativity should be divided, if at all, between Einstein and the mathematician David Hilbert.)

Those of you interested in how Einstein himself thought about special and general relativity might check out Einstein's own book, *The Meaning of Relativity*. Some of Einstein's formalism is old-fashioned (in particular, people don't use an imaginary time coordinate these days), and some of the notation and math will be beyond you, but you will likely find a good part of the book intelligible, and even where you can't follow the math you can often still get the gist of what's going on.

[15] Actually, definitely: you will see all the sights if you work through just two derivations: either §10.3.2 or §10.3.3, and either §10.3.4 or §10.3.7. In fact, if you could just work through either §10.3.4 or §10.3.7. But our inclusion of four derivations is not superfluous; each has a significantly different flavor.

figure, we have separated the red arrows for the pulse's trip out and its trip back so that you can see them distinctly; in reality they would of course lie along the same line.) If the perpendicular distance to the mirror is ℓ and the round-trip time is Δt, then the speed of the pulse is the round-trip distance 2ℓ divided by Δt: $c = 2\ell/\Delta t$. Thus we can express the round-trip time for the pulse as $\Delta t = 2\ell/c$.

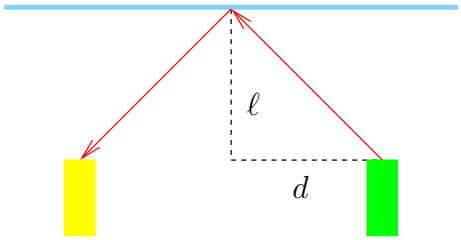


Figure 10.3: Light Pulse as Seen by an Observer Moving to the Right

If we now view the light pulse's trip to the mirror and back from the perspective of an observer traveling to the right at speed v, the laser, mirror, and light pulse, just like the phone poles by the side of the road when you are driving down the highway, all appear to be moving to the left at speed v. From the perspective of this observer, the pulse travels, not straight to and from the mirror, but along a triangular path, as shown by the red arrows in fig. (10.3). While the pulse is traveling to the mirror and back, the laser is traveling to the left at speed v from the perspective of the moving observer, so that the laser is at a different, more leftward location when the pulse returns than it was when the pulse was emitted; in fig. (10.3), the green rectangle represents the location of the laser when the pulse is emitted and the yellow rectangle represents the location of the laser when the pulse returns.[16] From the perspective of this moving frame, the perpendicular distance to the mirror is the same ℓ as in the "stationary" frame.[17] If we denote the time interval between the emission and reception of the pulse in this moving frame by $\Delta t'$,

[16]The mirror will of course also be moving to the left, but at the same speed v as everything else in the moving frame, so that the pulse will bounce off of the same part of the mirror that it did in the lab frame. Since what happens to the mirror isn't really material to the effects in which we are interested, we just made the mirror long in the figures in the hope that no one would even worry about it. But it seems that someone in the class always does worry about it, hence this footnote.

[17]Two issues: First, we put *stationary* in quotes because who is stationary and who is moving depends on whom you ask: to an observer in the moving frame, the "moving" frame is actually stationary, and it is the lab frame that appears to be moving — just as to a driver it is the car that seems to be at rest and the phone poles by the side of the road that seem to be moving. Hereafter, we will omit the quotation marks and expect you to be aware that *moving* and *stationary* are relative terms.

Second, you might think the assumption that the perpendicular distance ℓ is the same in the lab and moving frames unwarranted, and it is. Here we're just getting off the ground;

then by symmetry the time it takes the pulse to reach the mirror is $\frac{1}{2}\Delta t'$, and the sideways distance d traveled by the pulse on its way to the mirror is thus $d = v \cdot \frac{1}{2}\Delta t' = \frac{1}{2}v\Delta t'$. The distance covered by the pulse as it travels along the hypotenuse on its way to the mirror is therefore

$$\sqrt{\ell^2 + d^2} = \sqrt{\ell^2 + (\tfrac{1}{2}v\Delta t')^2}$$

and the total distance the pulse travels from laser to mirror and back to laser is just double this:

$$2\sqrt{\ell^2 + (\tfrac{1}{2}v\Delta t')^2}$$

If the speed of light, which would again be calculated by dividing the distance by the time, is to have the same value c for this moving observer as for everyone else, we must have

$$c = \frac{2\sqrt{\ell^2 + (\frac{1}{2}v\Delta t')^2}}{\Delta t'} = \sqrt{\left(\frac{2\ell}{\Delta t'}\right)^2 + v^2}$$

Solving this for $\Delta t'$, we obtain

$$\Delta t' = \frac{1}{\sqrt{1 - v^2/c^2}}\,\frac{2\ell}{c}$$

Recalling that the round-trip time in the stationary lab frame was $\Delta t = 2\ell/c$, we can write this as

$$\Delta t' = \frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t \tag{10.1}$$

Here we start to see some of the weird consequences of the postulate that the speed of light is the same in all reference frames: contrary to what your everyday common sense would lead you to expect, $\Delta t' \neq \Delta t$; the time interval between the emission and reception of the pulse is different for the two observers.

A similar weirdness occurs with the spatial displacements. Let us make the observers' axis of relative motion — what we have drawn as the horizontal direction in fig. (10.3) — our x axis, with the positive x direction to the right (that is, in the direction of motion of the moving observer). The perpendicular distance ℓ to the mirror is the same for both observers, but the displacement of the light pulse along the x axis is of course not the same because of their relative motion: in the stationary frame, the pulse returns to the same place it started out from, so $\Delta x = 0$ in that frame. But from the perspective of the moving observer, the laser and everything else appear

in a later, more sophisticated derivation of the Lorentz transform from symmetry this equality of perpendicular distances will be established rather than assumed (see p.548).

to be moving to the left (in the negative x direction) at speed v, so that the x displacement during the time interval $\Delta t'$ between the emission and reception of the pulse is

$$\Delta x' = -2d = -2 \cdot \tfrac{1}{2} v \Delta t' = -v \Delta t'$$

which, using our result (10.1) for $\Delta t'$, can be written

$$\Delta x' = -v \frac{1}{\sqrt{1 - v^2/c^2}} \Delta t = -\frac{1}{\sqrt{1 - v^2/c^2}} v \Delta t \tag{10.2}$$

In Newtonian physics, we would also have expected Δx and $\Delta x'$ to differ, but not in the same way: while Δx would again be zero, $\Delta x'$ would be simply $-v\Delta t$. The relativistic result for $\Delta x'$ differs from this by the same weird factor of $1/\sqrt{1 - v^2/c^2}$ that occurred in the result (10.1) for the time.

10.3.2 Derivation of the Lorentz Transform

More generally (that is, beyond the special case of a light pulse bouncing off of a mirror), we can derive what is known as the Lorentz transform. The Lorentz transform relates the spatial displacement and time interval between two events in one reference frame to the spatial displacement and time interval between those same two events in another reference frame. Conventionally, the stationary reference frame is referred to as the lab frame and the other reference frame as the moving frame.[18]

For concreteness, consider someone standing by the side of a straight road and another person driving down that road at a constant velocity v. For the two events, let's take, as the first, the driver takes a big swig of Coke,[19] and, as the second, the driver belches. We'll make the frame of the person standing by the roadside the lab frame and the frame of the driver the moving frame, with the road being the x axis. In general, we'll denote quantities in the lab frame without primes and quantities in the moving frame with primes. Based on our results for the mirror and light pulse, we expect that the spatial displacement and time interval between the swig and the belch will be different for the bystander and the driver: in our notation, the displacement along the x axis will be Δx and the time interval Δt for the roadside observer in the lab frame, while for the driver in the moving frame they will be $\Delta x'$ and $\Delta t'$.[20] Based on our results for the mirror and light

[18] Again, *stationary* and *moving* are relative terms; who is stationary and who is moving depends on whom you ask.

[19] Or Pepsi; it really doesn't make any difference.

[20] One can almost hear Zippy the Pinhead going "Chiastic structure! Chiastic structure! Chiastic structure!"

pulse, we expect that $\Delta x' \neq \Delta x$ and $\Delta t' \neq \Delta t$. What we want is to find the exact relation between $\Delta x'$ and $\Delta t'$ and Δx and Δt.

We will begin by assuming, as should seem plausible, that the relation we are seeking is linear, that is, that $\Delta x'$ and $\Delta t'$ are a linear combination of Δx and Δt:[21]

$$\begin{aligned} \Delta x' &= A\,\Delta x + B\,\Delta t \\ \Delta t' &= D\,\Delta x + E\,\Delta t \end{aligned} \tag{10.3}$$

where A, B, D, and E are constant coefficients — constant in the sense that, while they may (and we expect will) depend on the relative velocity of the two reference frames, they do not depend on the two events between which the Δ's are taken. Our task is to determine the values of these four coefficients, and for this purpose we can use our results from the special case of the mirror and light pulse. In that case, the emission and reception of the light pulse occurred at the same place in the stationary frame, so that $\Delta x = 0$. Using $\Delta x = 0$ and our results (10.1) and (10.2) in eqq. (10.3), we have

$$\begin{aligned} -\frac{1}{\sqrt{1 - v^2/c^2}}\, v\Delta t &= 0 + B\,\Delta t \\ \frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t &= 0 + E\,\Delta t \end{aligned}$$

which we can solve for B and E:

$$\begin{aligned} B &= -v\,\frac{1}{\sqrt{1 - v^2/c^2}} \\ E &= \frac{1}{\sqrt{1 - v^2/c^2}} \end{aligned} \tag{10.4}$$

With these values of B and E, eqq. (10.3) are reduced to

$$\Delta x' = A\,\Delta x - v\,\frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t \tag{10.5a}$$

$$\Delta t' = D\,\Delta x + \frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t \tag{10.5b}$$

To get a result for A, we return to the case of swigging and belching: since the driver, from his or her own perspective, is stationary, in the moving frame of the driver the swig and the belch occur at the same place. Thus the

[21] The linearity of Lorentz transforms is proved in Appendix D.

displacement between those two events is zero in the moving frame: $\Delta x' = 0$. But to the roadside observer in the lab frame, the driver is moving down the road at velocity v the whole time, so that between the swig and the belch the driver has moved a distance $\Delta x = v\Delta t$ down the x axis. Using $\Delta x' = 0$ and $\Delta x = v\Delta t$ in eq. (10.5a) gives

$$0 = A\, v\Delta t - v\,\frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t = \left(A - \frac{1}{\sqrt{1 - v^2/c^2}}\right) v\Delta t$$

which yields

$$A = \frac{1}{\sqrt{1 - v^2/c^2}}$$

so that eqq. (10.5) are further reduced to

$$\Delta x' = \frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta x - v\,\frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t \tag{10.6a}$$

$$\Delta t' = D\,\Delta x + \frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t \tag{10.6b}$$

Finally, to get a result for D, we leave the case of the swig and the belch and instead consider a light pulse traveling down the road. Since the speed of light must be the same for all observers, and in particular must be the same for the bystander and the driver, we have both

$$c = \frac{\Delta x}{\Delta t} \qquad \text{and} \qquad c = \frac{\Delta x'}{\Delta t'}$$

or, phrased the other way around,

$$\Delta x = c\Delta t \qquad \text{and} \qquad \Delta x' = c\Delta t' \tag{10.7}$$

Using these expressions for Δx and $\Delta x'$ in eq. (10.6a), we have

$$\begin{aligned} c\Delta t' &= \frac{1}{\sqrt{1 - v^2/c^2}}\,c\Delta t - v\,\frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t \\ &= \left(1 - \frac{v}{c}\right)\frac{1}{\sqrt{1 - v^2/c^2}}\,c\Delta t \end{aligned}$$

which, when we divide both sides by c, becomes

$$\Delta t' = \left(1 - \frac{v}{c}\right)\frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t \tag{10.8a}$$

And using relations (10.7) in eq. (10.6b), we also have

$$\Delta t' = D\, c\Delta t + \frac{1}{\sqrt{1 - v^2/c^2}}\, \Delta t \tag{10.8b}$$

Equating our two results for $\Delta t'$ ((10.8a) and (10.8b)) gives

$$\left(1 - \frac{v}{c}\right) \frac{1}{\sqrt{1 - v^2/c^2}}\, \Delta t = D\, c\Delta t + \frac{1}{\sqrt{1 - v^2/c^2}}\, \Delta t$$

which we can solve for D:

$$D = -\frac{v}{c^2}\, \frac{1}{\sqrt{1 - v^2/c^2}}$$

Using this result for D in eqq. (10.6), at long last we arrive at our result for the Lorentz transform:

$$\begin{aligned} \Delta x' &= \frac{1}{\sqrt{1 - v^2/c^2}}\, \Delta x - v \frac{1}{\sqrt{1 - v^2/c^2}}\, \Delta t \\ \Delta t' &= -\frac{v}{c^2} \frac{1}{\sqrt{1 - v^2/c^2}}\, \Delta x + \frac{1}{\sqrt{1 - v^2/c^2}}\, \Delta t \end{aligned}$$

Written this way, these relations are a bit unwieldy; we can simplify things by first pulling out the common factor of $1/\sqrt{1 - v^2/c^2}$,

$$\begin{aligned} \Delta x' &= \frac{1}{\sqrt{1 - v^2/c^2}} (\Delta x - v\Delta t) \\ \Delta t' &= \frac{1}{\sqrt{1 - v^2/c^2}} \left(-\frac{v}{c^2} \Delta x + \Delta t\right) \end{aligned}$$

If we then multiply the bottom equation through by c and at the same time re-express the v in the top equation as $c(v/c)$, we arrive at

$$\begin{aligned} \Delta x' &= \frac{1}{\sqrt{1 - v^2/c^2}} \left(\Delta x - \frac{v}{c}\, c\Delta t\right) \\ c\Delta t' &= \frac{1}{\sqrt{1 - v^2/c^2}} \left(-\frac{v}{c}\, \Delta x + c\Delta t\right) \end{aligned} \tag{10.9}$$

This way of writing the Lorentz transform is much simpler and more æsthetic.

10.3.3 A Nicer Derivation of the Lorentz Transform

In this section we will present a cleaner, more elegant derivation if the Lorentz transform.

With a great deal of déjà vu, we seek to relate the spatial displacement Δx and time interval Δt between two events according to an observer in the stationary frame to the spatial displacement $\Delta x'$ and time interval $\Delta t'$ between those same two events according to an observer in the moving frame, and we start by assuming that the relationship is linear: [22]

$$\begin{aligned} \Delta x' &= A\,\Delta x + B\,\Delta t \\ \Delta t' &= D\,\Delta x + E\,\Delta t \end{aligned} \tag{10.10}$$

Our task is to determine the values of the coefficients A, B, D, and E. These coefficients will be constants in the sense that, while they may (and we expect will) depend on the relative velocity of the two reference frames, they do not depend on the two events between which the Δ's are taken. We can therefore generate enough relations to determine these four unknown coefficients by applying eqq. (10.10) to simple special cases for which we know the Δx's and Δt's.

First consider the case of a light pulse moving in the positive x direction: since the speed of this light pulse must be the same in both reference frames, we must have both

$$c = \frac{\Delta x}{\Delta t} \qquad \text{and} \qquad c = \frac{\Delta x'}{\Delta t'}$$

or, phrased the other way around,

$$\Delta x = c\Delta t \qquad \text{and} \qquad \Delta x' = c\Delta t' \tag{10.11}$$

Using eqq. (10.11) in eqq. (10.10), we have

$$\begin{aligned} c\Delta t' &= A\,c\Delta t + B\,\Delta t \\ \Delta t' &= D\,c\Delta t + E\,\Delta t \end{aligned}$$

Dividing the former of these relations by the latter, we obtain

$$c = \frac{cA + B}{cD + E}$$

or in other words

$$cA + B = c^2 D + cE \tag{10.12}$$

This constitutes one of the four relations we need to solve for the coefficients A, B, D, and E.

[22] Again, the linearity of Lorentz transforms is proved in Appendix D.

We can generate a second relation among these coefficients simply by considering a light wave traveling in the negative x rather than the positive x direction: the chain of reasoning and calculations will be exactly the same as for the light wave traveling in the positive x direction, except that where before we had $+c$ we will now have $-c$. That is, we need simply replace all the c's in eq. (10.12) by $-c$, to obtain

$$-cA + B = c^2 D - cE \tag{10.13}$$

Taking the sum or difference of eqq. (10.12) and (10.13), we see that we must have

$$E = A \qquad\qquad B = c^2 D$$

Using these two results reduces eqq. (10.10) to

$$\begin{aligned} \Delta x' &= A\,\Delta x + c^2 D\,\Delta t \\ \Delta t' &= D\,\Delta x + A\,\Delta t \end{aligned} \tag{10.14}$$

To solve for the remaining coefficients A and D, consider the case of an object at rest in the moving frame. Since the object isn't going anywhere in the moving frame, $\Delta x' = 0$. From the perspective of the stationary frame, the object is moving at the velocity v of the moving frame, so that $\Delta x = v\Delta t$. Using $\Delta x' = 0$ and $\Delta x = v\Delta t$ in eqq. (10.14), we have

$$\begin{aligned} 0 &= A\,v\Delta t + c^2 D\,\Delta t \\ \Delta t' &= D\,v\Delta t + A\,\Delta t \end{aligned}$$

The former of these relations gives

$$D = -\frac{v}{c^2} A \tag{10.15a}$$

which, when substituted into the latter, yields

$$\Delta t' = \left(-\frac{v}{c^2} A\right) v\Delta t + A\,\Delta t = \left(1 - \frac{v^2}{c^2}\right) A\,\Delta t \tag{10.15b}$$

Since we don't have a relation between Δt and $\Delta t'$ in terms of known quantities, this last relation doesn't look like it's going to help us.

To get the remaining relation that we need to determine A and D, we will now use the same sort of lazy trick that we did above with the light wave: instead of an object at rest in the moving frame, we will consider an object at rest in the stationary frame. This time the object isn't going anywhere in the stationary frame, so that $\Delta x = 0$. And from the perspective of the moving frame, the object is, like the phone poles by the side of the road,

moving backward with velocity $-v$, so that $\Delta x' = -v\Delta t'$. Using $\Delta x = 0$ and $\Delta x' = -v\Delta t'$ in eqq. (10.14), we have

$$-v\Delta t' = c^2 D\, \Delta t \tag{10.16a}$$

$$\Delta t' = A\, \Delta t \tag{10.16b}$$

If we divide the former of these relations by the latter, we obtain

$$-v = \frac{c^2 D}{A}$$

or

$$D = -\frac{v}{c^2} A$$

which is redundant with eq. (10.15a) and therefore not going to help us. To the faint of heart it might seem that all is lost and that life is just a cruel joke after all. But wait: eq. (10.16b) relates the time Δt in the frame of the object to the time $\Delta t'$ in a frame moving at velocity v relative to the object. And eq. (10.15b) is the same relation, just with the frames reversed: in eq. (10.15b), the object is in the moving frame, from the perspective of which the stationary frame is moving backward with velocity $-v$. Eq. (10.15b) therefore relates the time $\Delta t'$ in the frame of the object to the time Δt in a frame moving with velocity $-v$ relative to the object. But which direction we make the positive x direction is arbitrary; it should not make any physical difference whether we look at the time in a frame moving at $+v$ relative to the object or a frame moving at $-v$ relative to the object. Eqq. (10.16b) and (10.15b) are therefore the same equation, just with their Δt's and $\Delta t'$'s reversed: all we have to do to "equate" these two relations is interchange the Δt and $\Delta t'$ in one of them. If we interchange the Δt and $\Delta t'$ in eq. (10.16b), we therefore conclude that [23]

$$\Delta t' = \left(1 - \frac{v^2}{c^2}\right) A\, \Delta t \qquad \text{must be equivalent to} \qquad \Delta t = A\, \Delta t'$$

To solve for A, we just use the latter relation in the former:

$$\Delta t' = \left(1 - \frac{v^2}{c^2}\right) A(A\, \Delta t') = \left(1 - \frac{v^2}{c^2}\right) A^2 \Delta t'$$

[23] An equivalent but different way to look at things: Eq. (10.16b) transforms the time interval as we go into a frame moving at relative velocity $+v$; eq. (10.15b) transforms the time interval as we go into a frame moving at relative velocity $-v$. These two transforms, done in succession, should therefore bring us back to our original frame and original time interval. In other words, the factors in the transforms (10.16b) and (10.15b), when compounded, should yield the identity:

$$A\left(1 - \frac{v^2}{c^2}\right) A = 1$$

which likewise yields $A = 1/\sqrt{1 - v^2/c^2}$.

which yields

$$A = \frac{1}{\sqrt{1 - v^2/c^2}} \tag{10.17a}$$

Eq. (10.15a) then gives

$$D = -\frac{v}{c^2}A = -\frac{v}{c^2}\frac{1}{\sqrt{1 - v^2/c^2}} \tag{10.17b}$$

With eqq. (10.17), our Lorentz transform (10.14) becomes

$$\begin{aligned}
\Delta x' &= A\,\Delta x + c^2 D\,\Delta t \\
&= \frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta x + c^2\left(-\frac{v}{c^2}\frac{1}{\sqrt{1 - v^2/c^2}}\right)\Delta t \\
&= \frac{1}{\sqrt{1 - v^2/c^2}}(\Delta x - v\Delta t) \\
\Delta t' &= D\,\Delta x + A\,\Delta t \\
&= -\frac{v}{c^2}\frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta x + \frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t \\
&= \frac{1}{\sqrt{1 - v^2/c^2}}\left(-\frac{v}{c^2}\Delta x + \Delta t\right)
\end{aligned}$$

If we make this a bit more æsthetic by multiplying the second relation through by c and re-expressing the v in the first relation as $c(v/c)$, our result for the Lorentz transform takes the same form that we obtained in the previous section (eqq. (10.9)):

$$\begin{aligned}
\Delta x' &= \frac{1}{\sqrt{1 - v^2/c^2}}\left(\Delta x - \frac{v}{c}\,c\Delta t\right) \\
c\Delta t' &= \frac{1}{\sqrt{1 - v^2/c^2}}\left(-\frac{v}{c}\,\Delta x + c\Delta t\right)
\end{aligned} \tag{10.18}$$

10.3.4 A More Modern Derivation of the Lorentz Transform

In this section we will present yet another derivation of the Lorentz transform, this one more in keeping with the modern understanding of the theory.

Though it may feel like the movie *Groundhog Day*, once again we seek to relate the spatial displacement Δx and time interval Δt between two events according to an observer in the stationary frame to the spatial displacement

$\Delta x'$ and time interval $\Delta t'$ between those same two events according to an observer in the moving frame, and we start by assuming that the relationship is linear: [24]

$$\begin{aligned} \Delta x' &= A\,\Delta x + B\,\Delta t \\ \Delta t' &= D\,\Delta x + E\,\Delta t \end{aligned} \tag{10.19}$$

Our task is to determine the values of the coefficients A, B, D, and E — coefficients we expect will depend on the relative velocity of the two reference frames but not on the two events between which the Δ's are taken.

Consider the case of a light pulse moving in the positive x direction. In the stationary frame, the pulse moves forward through a spatial displacement Δx during a time interval Δt, with the speed c of the pulse given as usual by displacement divided by time:

$$c = \frac{\Delta x}{\Delta t} \tag{10.20}$$

In the moving frame, the corresponding spatial displacement and time interval are $\Delta x'$ and $\Delta t'$, and although these will, we expect, differ from Δx and Δt, in both frames the light pulse will be moving at the same speed c, so that we also have

$$c = \frac{\Delta x'}{\Delta t'} \tag{10.21}$$

The next step in the reasoning — a critically important one — is to recognize that relativistic effects are a consequence, not of the properties of light, but of the properties of spacetime. So while eqq. (10.20) and (10.21) apply very specifically to the progress of a light pulse, the Lorentz transform should be much more general; it should apply to the spatial displacement and time interval between any two events, not just those associated with the progress of a light pulse. We are thus led to try to rewrite eqq. (10.20) and (10.21) in a more general form that can be applied to any Δx and Δt. If we gerrymander these equations by squaring both sides, multiplying both sides by Δt^2 (or $\Delta t'^2$), and moving everything to the left side of the equation, we end up with the vaguely Pythagorean-looking relations [25]

$$c^2\Delta t^2 - \Delta x^2 = 0 \qquad\qquad c^2\Delta t'^2 - \Delta x'^2 = 0 \tag{10.22}$$

Eqq. (10.22) tell us two things: first, that for light the combination $c^2\Delta t^2 - \Delta x^2$ is invariant (that is, it has the same value for all observers); and second, that for light the invariant value of this combination happens to be zero.

[24] The linearity of Lorentz transforms is proved in Appendix D.

[25] This similarity between eqq. (10.22) and familiar Pythagorean combinations like $\Delta x^2 + \Delta y^2$ is more than coincidental and is in fact the most compelling reason for writing eqq. (10.20) and (10.21) in the form (10.22). The connection may well seem a little tenuous from our current vantage point, but it will become more clear in §10.3.7.

Now, even if you were a bit doubtful or skeptical of our gerrymandering of eqq. (10.20) and (10.21) into eqq. (10.22), there is no question that eqq. (10.22) hold, and you have to admit that it would be very odd if the combination $c^2\Delta t^2 - \Delta x^2$ were invariant only for light: this would mean that space and time (Δx and Δt) had some special property that held only for light. But why should such a property apply only to light? If it is truly a property of spacetime, it should apply to everything, since all objects, not just light, exist in and travel through spacetime. We are thus led to conclude that the combination $c^2\Delta t^2 - \Delta x^2$ should be invariant for all objects and indeed for the Δx and Δt between any two events. This combination may be zero only for light,[26] but whatever its value for other objects, the combination should nonetheless be invariant for them. We therefore generalize eqq. (10.22) to

$$c^2\Delta t'^2 - \Delta x'^2 = c^2\Delta t^2 - \Delta x^2 \tag{10.23}$$

Eq. (10.23) is simply stating explicitly that the combination $c^2\Delta t^2 - \Delta x^2$ has the same value in all reference frames, regardless of the nature of the two events to which the Δx's and Δt's refer.[27]

If now we substitute the general relationships (10.19) into eq. (10.23), we have

$$c^2(D\,\Delta x + E\,\Delta t)^2 - (A\,\Delta x + B\,\Delta t)^2 = c^2\Delta t^2 - \Delta x^2$$

which, when we do out the squares and recombine terms, becomes

$$\begin{aligned}(c^2E^2 - B^2)\,\Delta t^2 + (c^2DE - AB)\,2\Delta t\Delta x \\ +\ (c^2D^2 - A^2)\,\Delta x^2 = c^2\Delta t^2 - \Delta x^2\end{aligned} \tag{10.24}$$

Now, since eq. (10.24) must hold for any two events, it must hold for all possible Δx and Δt. Since we can vary Δx and Δt independently, we can separately equate the Δt^2, $2\Delta t\Delta x$, and Δx^2 terms in eq. (10.24):

$$c^2E^2 - B^2 = c^2 \tag{10.25a}$$

$$c^2DE - AB = 0 \tag{10.25b}$$

$$c^2D^2 - A^2 = -1 \tag{10.25c}$$

We thus have three equations in the four unknown coefficients A, B, D, and E. To obtain a fourth equation that will enable us to solve for these

[26] More strictly, it is zero only for those objects that move at the speed of light, which includes gravitons and other massless types of matter in addition to light.

[27] As we will see in §10.5, $c^2\Delta t^2 - \Delta x^2$ is the square of what is known as the invariant interval and is related to the time elapsed in an object's own reference frame, which is known as the proper time. In fact, you can see that right now: in its own reference frame, an object is at rest, so that $\Delta x = 0$ and $c^2\Delta t^2 - \Delta x^2 = c^2\Delta t^2$, where Δt is the time elapsed in the object's frame, that is, on the object's own clock.

coefficients, it will suffice to consider a simple case for which we know the relationship among Δx, Δt, $\Delta x'$ and $\Delta t'$: the case of an object at remaining at rest in the stationary frame F. Because the object isn't moving in frame F, $\Delta x = 0$. And if the frame F' is moving at velocity v relative to frame F, then in frame F' the object will, like the phone poles by the side of the road, appear to be moving backward, with velocity $-v$. After a time interval $\Delta t'$ elapses in frame F' (corresponding to a time interval Δt elapsing in frame F), the object's displacement in frame F' will therefore be $\Delta x' = -v\Delta t'$. Using $\Delta x = 0$ and $\Delta x' = -v\Delta t'$ in eqq. (10.19), we have

$$\begin{aligned} -v\Delta t' &= B\,\Delta t \\ \Delta t' &= E\,\Delta t \end{aligned}$$

If we divide the former of these relations by the latter, we arrive at the fourth relation we need among A, B, D, and E:

$$\frac{B}{E} = -v \qquad \text{or} \qquad B = -vE \tag{10.25d}$$

Everything now boils down to pure algebra. Using eq. (10.25d) in eq. (10.25a), we have

$$\begin{aligned} c^2 &= c^2E^2 - B^2 \\ &= c^2E^2 - (-vE)^2 \\ &= c^2\left(1 - \frac{v^2}{c^2}\right)E^2 \end{aligned}$$

which yields

$$E = \frac{1}{\sqrt{1 - v^2/c^2}} \tag{10.26a}$$

and hence, from eq. (10.25d),

$$B = -vE = -v\frac{1}{\sqrt{1 - v^2/c^2}} \tag{10.26b}$$

From eq. (10.25b), we have

$$\begin{aligned} 0 &= c^2DE - AB \\ c^2DE &= AB \end{aligned}$$

which, with eq. (10.25d), gives

$$D = \frac{1}{c^2}A\frac{B}{E} = \frac{1}{c^2}A(-v) = -\frac{v}{c^2}A \tag{10.26c}$$

Using this result in eq. (10.25c), we then have

$$-1 = c^2D^2 - A^2 = c^2\left(-\frac{v}{c^2}A\right)^2 - A^2 = -\left(1 - \frac{v^2}{c^2}\right)A^2$$

so that

$$A = \frac{1}{\sqrt{1 - v^2/c^2}} \tag{10.26d}$$

$$D = -\frac{v}{c^2}A = -\frac{v}{c^2}\frac{1}{\sqrt{1 - v^2/c^2}} \tag{10.26e}$$

When our solutions (10.26b) through (10.26e) are used for A, B, D, and E in eq. (10.19), we arrive at the same result (10.9) for the Lorentz transform as in the preceding two sections:

$$\begin{aligned}
\Delta x' &= A\,\Delta x + B\,\Delta t \\
&= \frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta x - v\frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t \\
&= \frac{1}{\sqrt{1 - v^2/c^2}}\left(\Delta x - \frac{v}{c}\,c\Delta t\right) \\
c\Delta t' &= D\,\Delta x + E\,\Delta t \\
&= -\frac{v}{c^2}\frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta x + \frac{1}{\sqrt{1 - v^2/c^2}}\,\Delta t \\
&= \frac{1}{\sqrt{1 - v^2/c^2}}\left(-\frac{v}{c}\,\Delta x + c\Delta t\right)
\end{aligned}$$

10.3.5 Some Observations & Notation

Now that we've derived the Lorentz transform (10.9), let's be very clear about what it means.

$$\Delta x' = \frac{1}{\sqrt{1 - v^2/c^2}}\left(\Delta x - \frac{v}{c}\,c\Delta t\right)$$

$$c\Delta t' = \frac{1}{\sqrt{1 - v^2/c^2}}\left(-\frac{v}{c}\,\Delta x + c\Delta t\right)$$

We have two reference frames: the (unprimed) lab frame, which we regard as stationary, and the (primed) moving frame, which is traveling along the x axis at velocity v relative to the lab frame. We also have two distinct events, say

a and b. The $\Delta x = x_b - x_a$ and $\Delta t = t_b - t_a$ are the spatial displacement and time interval between these two events *in the lab frame*; the $\Delta x' = x'_b - x'_a$ and $\Delta t' = t'_b - t'_a$ are the spatial displacement and time interval between these same two events *in the moving frame*. The Lorentz transform gives us relations between the spatial displacement and time interval between two events in the lab frame and the corresponding spatial displacement and time interval between the same two events in the moving frame. If, for example, a driver takes a swig of a Coke and then later, at a point farther down the road, belches, the Lorentz transform allows us to relate the time and distance between the swig and the belch according to a roadside bystander to the time and distance between the swig and the belch according to the driver.

In these Lorentz relations we can see explicitly some of the weirdness that happens relativistically to spatial distances and time intervals: in both equations there is the strange factor $1/\sqrt{1 - v^2/c^2}$, and from the bottom equation we see that in general the time intervals $\Delta t'$ and Δt are not equal. In the sections that follow, we will work out some of the more fundamental manifestations and consequences of this weirdness. We will, however, make one very general and very fundamental observation here: for ordinary, everyday velocities, which are *far* less than the speed of light, the factor $1/\sqrt{1 - v^2/c^2}$ is very nearly just 1, and the Lorentz transform then becomes

$$\begin{aligned}
\Delta x' &= \frac{1}{\sqrt{1 - v^2/c^2}} \left(\Delta x - \frac{v}{c}\, c\Delta t \right) \\
&\approx 1 \cdot (\Delta x - v\Delta t) && \text{for } v \ll c \\
&= \Delta x - v\Delta t \\
c\Delta t' &= \frac{1}{\sqrt{1 - v^2/c^2}} \left(-\frac{v}{c}\, \Delta x + c\Delta t \right) \\
&\approx 1 \cdot (-0 \cdot \Delta x + c\Delta t) && \text{for } v \ll c \\
&= c\Delta t
\end{aligned}$$

Thus for ordinary, everyday velocities, the Lorentz transform just gives us the familiar Newtonian results: the time intervals are the same for both observers, and the spatial displacements differ only by the distance $v\Delta t$ that the moving observer travels during the time interval Δt.

The limit of velocities far less than the speed of light ($v \ll c$) is called the nonrelativistic limit.[28] In this limit, relativistic results should (and in fact do) always turn out to reproduce the familiar, common-sense results we get from Newtonian physics. Although in principle relativistic effects are always present, it is only when velocities comparable to the speed of light

[28] The other extreme, the limit of velocities very close to the speed of light, is called the ultrarelativistic limit. Just in case you were curious.

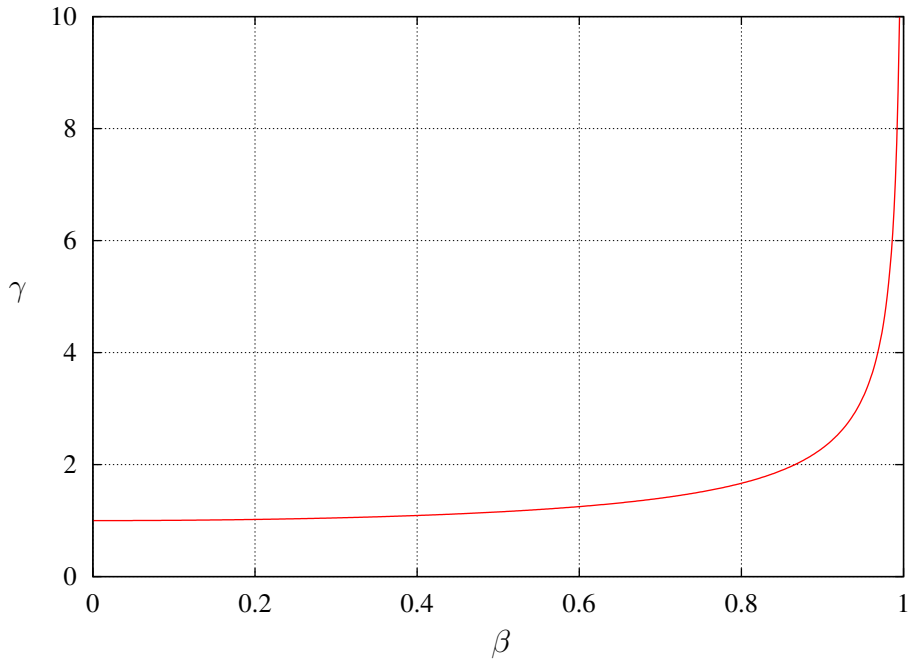


Figure 10.4: $\gamma = \dfrac{1}{\sqrt{1-\beta^2}}$ for $0 \leq \beta < 1$

are involved that we get quantitatively significant relativistic effects.[29] This in fact explains how relativity can be correct even though its consequences often contradict our common sense: our common sense is based entirely on velocities far less than the speed of light, for which the relativistic effects are imperceptibly small.

Two expressions involving the velocity v occur very frequently in relativistic calculations: v/c and $1/\sqrt{1-v^2/c^2}$. There is therefore a special notation for these combinations:

$$\beta = \frac{v}{c} \tag{10.27}$$

is the fraction of the speed of light at which the object is moving, and

$$\gamma = \frac{1}{\sqrt{1-v^2/c^2}} = \frac{1}{\sqrt{1-\beta^2}} \tag{10.28}$$

Sometimes γ is whimsically called the "warp factor," but neither β nor γ has a special name that is widely used; they are simply called *beta* and *gamma*. Note that as we go from rest up to the speed of light, $\beta = v/c$ goes from 0 to 1 and γ from 1 to ∞.[30] In particular, $\gamma \geq 1$ always. Fig. (10.4) shows a plot of γ from $\beta = 0$ to $\beta = 1$.

[29] It is a common error to suppose that there are *no* relativistic effects for velocities much less than the speed of light: there are always relativistic effects; it is just that they are imperceptibly small for velocities much less than the speed of light.

[30] As we will see later, it is in fact this divergence of γ as $v \to c$ that makes it impossible for massive objects ever to reach the speed of light.

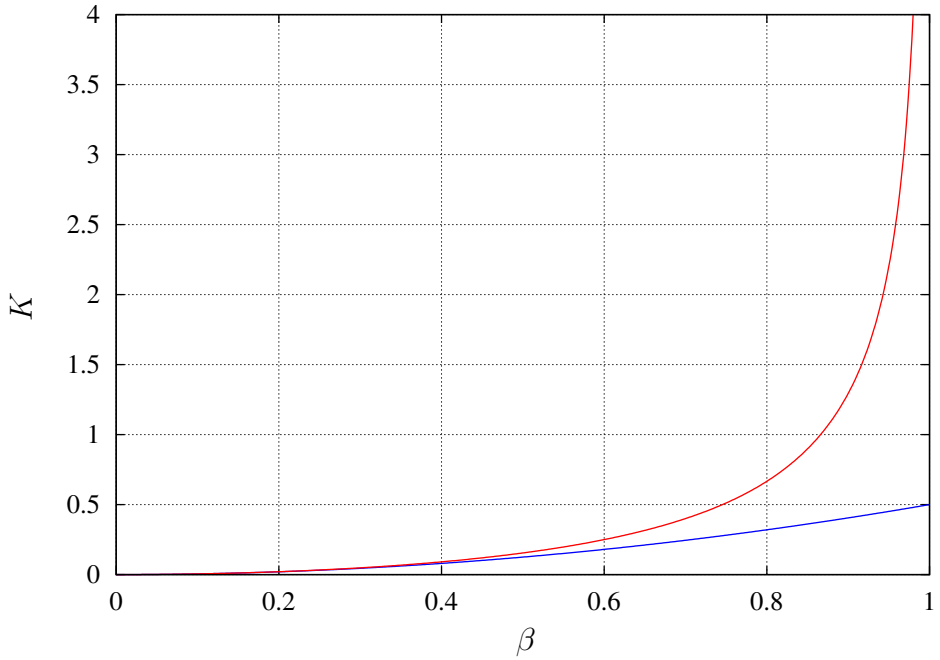


Figure 10.5: Relativistic versus Newtonian Kinetic Energy

As an example of how relativistic results agree to a very good approximation with familiar Newtonian results in the limit of low velocities, but differ markedly near the speed of light, fig. (10.5) shows the plots of the kinetic energy of a mass m for speeds from zero up to the speed of light: the blue line is the familiar Newtonian $\frac{1}{2}mv^2$, the red line the relativistic kinetic energy (which, as we will see in §10.7, is $(\gamma - 1)mc^2$). The values on the vertical axis are in units of mc^2 (so that, for example, at $v = c$ the value of .5 for the Newtonian kinetic energy corresponds to $\frac{1}{2}mc^2/mc^2$).

In terms of β and γ, the Lorentz transform (10.9) becomes

$$\Delta x' = \gamma(\Delta x - \beta c\Delta t) \tag{10.29a}$$
$$c\Delta t' = \gamma(-\beta\Delta x + c\Delta t) \tag{10.29b}$$

which is obviously much simpler and more æsthetic. In fact, if you are familiar with matrix multiplication, the Lorentz transform can be written even more simply and æsthetically as the single matrix equation

$$\begin{pmatrix} \Delta x' \\ c\Delta t' \end{pmatrix} = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} \Delta x \\ c\Delta t \end{pmatrix} \tag{10.30}$$

10.3.6 The Inverse Lorentz Transform

If we go from our reference frame into a frame moving at velocity v relative to us and then, from that frame, into a frame moving at relative velocity $-v$, we should find ourselves back in our original reference frame. We therefore expect that the inverse Lorentz transform is simply the Lorentz transform in which the frame F' is moving at velocity $-v$ relative to the frame F. In other words, if we denote by $\Lambda(v)$ the Lorentz transform that takes us from a frame F into a frame F' moving at velocity v relative to F, then the inverse Lorentz transform should be $\Lambda(-v)$.

We will verify this first using the matrix form of the Lorentz transform and then, for the those of you who aren't familiar with matrices, establish this same result without matrices. Either way, first note that if we make the change $v \to -v$, then

$$\begin{aligned} \beta &= \frac{v}{c} \to -\frac{v}{c} = -\beta \\ \gamma &= \frac{1}{\sqrt{1-\beta^2}} \to \frac{1}{\sqrt{1-(-\beta)^2}} = \frac{1}{\sqrt{1-\beta^2}} = \gamma \end{aligned} \tag{10.31}$$

That is, we should everywhere replace β by $-\beta$, but leave γ unchanged.

Now, then, first with matrices: The inverse of the matrix form of the Lorentz transform (10.30),

$$\Lambda(v) = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \tag{10.32}$$

should therefore be

$$\Lambda^{-1}(v) = \Lambda(-v) = \gamma \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix} \tag{10.33}$$

Then we will have

$$\begin{aligned} \Lambda^{-1}(v)\Lambda(v) &= \gamma \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix} \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \\ &= \gamma^2 \begin{pmatrix} 1 & \beta \\ \beta & 1 \end{pmatrix} \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \\ &= \gamma^2 \begin{pmatrix} 1-\beta^2 & 0 \\ 0 & 1-\beta^2 \end{pmatrix} \\ &= \left(\frac{1}{\sqrt{1-\beta^2}} \right)^2 \begin{pmatrix} 1-\beta^2 & 0 \\ 0 & 1-\beta^2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{aligned}$$

which is exactly what we want: $\Lambda^{-1}(v)\Lambda(v)$ is the identity matrix, that is, is the identity transform, which doesn't change anything. And as you can easily check for yourself, we get the same result for $\Lambda(v)\Lambda^{-1}(v)$.

For those of you unfamiliar with matrices, we can verify that $\Lambda(-v)$ is the inverse of $\Lambda(v)$ the long way, by first going from Δx and Δt in a frame F to $\Delta x'$ and $\Delta t'$ in a frame F' moving at velocity v relative to F and from there to $\Delta x''$ and $\Delta t''$ in a frame F'' moving at velocity $-v$ relative to F': from eqq. (10.29) and (10.31) we have

$$\Delta x' = \gamma(\Delta x - \beta c\Delta t) \qquad c\Delta t' = \gamma(-\beta\Delta x + c\Delta t) \tag{10.34}$$

and then

$$\Delta x'' = \gamma(\Delta x' + \beta c\Delta t') \qquad c\Delta t'' = \gamma(+\beta\Delta x' + c\Delta t') \tag{10.35}$$

so that, using eqq. (10.34) in eqq. (10.35),

$$\begin{aligned}
\Delta x'' &= \gamma(\Delta x' + \beta c\Delta t') \\
&= \gamma\Big(\big(\gamma(\Delta x - \beta c\Delta t)\big) + \beta\big(\gamma(-\beta\Delta x + c\Delta t)\big)\Big) \\
&= \gamma^2(1-\beta^2)\Delta x \\
&= \left(\frac{1}{\sqrt{1-\beta^2}}\right)^2(1-\beta^2)\Delta x \\
&= \Delta x \\
c\Delta t'' &= \gamma(+\beta\Delta x' + c\Delta t') \\
&= \gamma\Big(\beta\big(\gamma(\Delta x - \beta c\Delta t)\big) + \big(\gamma(-\beta\Delta x + c\Delta t)\big)\Big) \\
&= \gamma^2(1-\beta^2)c\Delta t \\
&= c\Delta t
\end{aligned}$$

Thus $\Delta x''$ and $\Delta t''$ are the same as Δx and Δt for all Δx and Δt, which establishes that $\Lambda(-v)$ does indeed "undo" $\Lambda(v)$.

The inverse Lorentz transform is thus

$$\Delta x' = \gamma(\Delta x + \beta c\Delta t) \tag{10.36a}$$
$$c\Delta t' = \gamma(\beta\Delta x + c\Delta t) \tag{10.36b}$$

or, in matrix form,

$$\begin{pmatrix}\Delta x' \\ c\Delta t'\end{pmatrix} = \gamma\begin{pmatrix}1 & \beta \\ \beta & 1\end{pmatrix}\begin{pmatrix}\Delta x \\ c\Delta t\end{pmatrix} \tag{10.37}$$

10.3.7 The Lorentz Transform from Symmetry

In this section, we present yet another derivation of the Lorentz transform, this time working out the consequences of spacetime symmetry in a more sophisticated way that, while admittedly a bit more involved than strictly necessary for our present purposes, is the most modern way of deriving the Lorentz transform and formulating the theory of relativity. If you go on in physics, you will be using these kinds of techniques and reasoning all the time. It's the big-people way to do things.

Symmetries consist of two fundamental parts: a symmetry operation and a quantity that is invariant under the symmetry operation. A circle, for example, has a geometric symmetry: the circle is symmetric under rotations about its center. In this case, the symmetry operation is the rotation and the invariant is the circle itself. A square is similarly invariant under rotations by multiples of 90°. More abstract, nongeometric symmetries are of course possible: for an income indexed to inflation (that is, an income that increases at the same rate as inflation), the symmetry operation is inflation and the invariant is the buying power of that income.

In relativity, the reasoning from symmetry starts out like that of §10.3.4: if Δx and Δt represent the displacement and corresponding elapsed time for a light pulse traveling along the positive x axis, then the speed c of the light pulse must be given by

$$c = \frac{\Delta x}{\Delta t} \tag{10.38}$$

The corresponding spatial displacement and time interval $\Delta x'$ and $\Delta t'$ in a different reference frame will, we expect, differ from Δx and Δt, but in both frames the light pulse will be moving at the same speed c, so that we also have

$$c = \frac{\Delta x'}{\Delta t'} \tag{10.39}$$

The next step in the reasoning — a critically important one — is to recognize that relativistic effects are a consequence, not of the properties of light, but of the properties of spacetime. So while eqq. (10.38) and (10.39) apply very specifically to the progress of a light pulse, the Lorentz transform should be much more general; it should apply to the spatial displacement and time interval between any two events, not just those associated with the progress of a light pulse. We are thus led to try to rewrite eqq. (10.38) and (10.39) in a more general form that can be applied to any Δx and Δt. If we gerrymander these equations by squaring both sides, multiplying both sides by Δt^2 (or $\Delta t'^2$), and moving everything to the left side of the equation, we end up with

the vaguely Pythagorean-looking relations[31]

$$c^2\Delta t^2 - \Delta x^2 = 0 \qquad\qquad c^2\Delta t'^2 - \Delta x'^2 = 0 \tag{10.40}$$

Eqq. (10.40) and tell us two things: first, that for light the combination $c^2\Delta t^2 - \Delta x^2$ is invariant (in the sense that it has the same value for all observers); and second, that for light the invariant value of this combination happens to be zero.

Even if our gerrymandering of eqq. (10.20) and (10.21) into eqq. (10.40) seemed a bit dubious, eqq. (10.40) certainly hold, and it would be very odd if the combination $c^2\Delta t^2 - \Delta x^2$ were invariant only for light, because this would mean that spacetime (Δx and Δt) had some special property that held only for light. We are thus led to postulate that the invariance of $c^2\Delta t^2 - \Delta x^2$ is a property of spacetime and should therefore apply universally, not just to light: the value of $c^2\Delta t^2 - \Delta x^2$ will in general not be zero for the Δx and Δt between two events unrelated to the progress of a light pulse — in particular, it will not be zero for the Δx and Δt on the trajectory of an object moving at a speed other than that of light — , but whatever its value, that value should be the same for all observers. We are therefore led to generalize eqq. (10.40) to

$$c^2\Delta t'^2 - \Delta x'^2 = c^2\Delta t^2 - \Delta x^2 \tag{10.41}$$

In other words, nature has a symmetry under which $c^2\Delta t^2 - \Delta x^2$ is invariant. Our task is now to determine the most general form of the symmetry operation that leaves the quantity $c^2\Delta t^2 - \Delta x^2$ invariant; it is this symmetry operation that will constitute the Lorentz transform.

We begin by setting up our two reference frames: frame F, with coordinates (x, t), and, moving at velocity v relative to F, frame F', with coordinates (x', t'). Without loss of generality, we can set up our x, t, x', and t' axes so that F and F' to have a common spacetime origin, that is, so that the spacetime point $(x, t) = (0, 0)$ is the same as $(x', t') = (0, 0)$. If we choose this common origin as the "initial" point for the differences denoted by our

[31] With just an x axis, you might question why we square both sides rather than writing this relation simply as $c\Delta t - \Delta x = 0$. Consider, however, a light pulse moving in a general three-dimensional direction, so that it has displacements Δx, Δy, and Δz: in this case the distance traveled by pulse is $\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}$, and

$$c^2\Delta t^2 - (\Delta x^2 + \Delta y^2 + \Delta z^2) = 0$$

is much nicer and more tractable mathematically than the ugly

$$c\Delta t - \sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2} = 0$$

And there are also more sophisticated reasons for using the squares, related to differential geometry and general relativity (see §10.9).

Δ's, then eq. (10.41) becomes

$$c^2(t'-0)^2-(x'-0)^2=c^2(t-0)^2-(x-0)^2$$
$$c^2t'^2-x'^2=c^2t^2-x^2 \tag{10.42}$$

We want to determine the most general relation

$$\begin{aligned} x' &= x'(x,t) \\ t' &= t'(x,t) \end{aligned} \tag{10.43}$$

for which eq. (10.42) holds. This would otherwise be daunting, but the magnitude of the task is greatly reduced by the proof, in Appendix D, that Lorentz transforms must be linear.[32] As will become clear shortly, it will also help to consider the special case of an infinitesimal relative velocity v between the frames F and F': infinitesimal transforms are easier to tackle directly than finite transforms, and one can then obtain results for finite transforms by integrating over a succession of infinitesimal transforms. In our present context of Lorentz transforms, note that if the relative velocity between the two frames vanished, there would be no difference between the coordinates x and t and the coordinates x' and t': F and F' would be the same frame, so that we would of course have $x'=x$ and $t'=t$. With an infinitesimal relative velocity between the frames, the differences between the coordinates x and t and the coordinates x' and t' will be correspondingly infinitesimal. Let us denote these differences by dx and dt:

$$\begin{aligned} c\,dt &= ct'-ct \\ dx &= x'-x \end{aligned} \tag{10.44}$$

Since Lorentz transforms are linear, for an infinitesimal relative velocity between F and F' the dx and dt of eqq. (10.44) must be linear combinations of x and t:[33]

$$\begin{aligned} c\,dt &= A\,d\alpha\,ct + B\,d\alpha\,x \\ dx &= D\,d\alpha\,ct + E\,d\alpha\,x \end{aligned} \tag{10.45}$$

[32] It would be really nice to include this proof here for the sake of completeness, but to do so in-line, so to speak, would distract us too much from the aspects of the derivation that are our principal concern at present.

[33] While everything we do here is correct, there are a couple of things we should note. First, by having stipulated that F and F' have a common origin in spacetime, our derivation overlooks the possibility spacetime translations, that is, shifts by constant amounts along our x and t axes. But while such translations also leave the interval invariant, they have nothing to do with the relative velocity between reference frames and are therefore not considered to be Lorentz transforms. Rather, spacetime translations together with the Lorentz transform comprise what are known as the Poincaré group.

Second, while Lorentz transforms and translations are the only continuous transforms that leave the interval invariant, there are additional discrete transforms that do so: since the interval depends on the *squares* of time intervals and spatial displacements, changes of sign won't matter, and therefore a parity operation that inverts the spatial axes ($(x,y,z)\to(-x,-y,-z)$) or a time reversal operation ($t\to -t$) will also leave the interval invariant. But we are not concerned here with these sorts of transforms, either.

where A, B, C, and D are as yet unknown coefficients and α is some parameter that is small when the relative velocity between frames is small and that we expect will turn out to be a function of the velocity v of F' relative to F. The coefficients A, B, C, and D are constant in the sense that they do not depend on x or t, but they may turn out to be functions of α. And we have written $d\alpha$ rather than just α in eqq. (10.45) to make explicit that the contributions of x and t to dx and dt are infinitesimal. Our task is now to determine α and the coefficients A, B, C, and D. To accomplish this, we re-express eqq. (10.44) as

$$\begin{aligned} ct' &= ct + c\,dt \\ x' &= x + dx \end{aligned}$$

and impose the invariance condition (10.42):

$$\begin{aligned} c^2t'^2 - x'^2 &= c^2t^2 - x^2 \\ (ct + c\,dt)^2 - (x + dx)^2 &= c^2t^2 - x^2 \end{aligned}$$

which, when we expand the squares, becomes

$$c^2t^2 + 2ct(c\,dt) + (c\,dt)^2 - x^2 + 2x\,dx + dx^2 = c^2t^2 - x^2$$

In terms of infinitesimal quantities, there are three orders of terms in this relation: zero-order (noninfinitesimal) terms that go as dx^0 and dt^0, first-order terms that go as dt^1 or dx^1, and second-order terms that go as dx^2 or dt^2. The zero-order terms (the c^2t^2 and the x^2) cancel out. Of the remaining terms, we may neglect the second-order terms in comparison with those that are only first-order, since these second-order correction terms are infinitesimally smaller than the leading first-order terms. We are thus left with the condition

$$2ct(c\,dt) - 2x\,dx = 0 \tag{10.46}$$

Using eqq. (10.45) in eq. (10.46), we have

$$2ct(A\,d\alpha\,ct + B\,d\alpha\,x) - 2x(D\,d\alpha\,ct + E\,d\alpha\,x) = 0$$

which, when we expand and divide both sides by $2\,d\alpha$, becomes

$$Ac^2t^2 + (B - D)x\,ct - Ex^2 = 0$$

Now, this equation must hold for *all* x and t, and since we can vary x and t independently of each other, this means that the equality must hold separately for the terms in t^2, the terms in $x\,ct$, and the terms in x^2:

$$Ac^2t^2 = 0 \qquad (B - D)x\,ct = 0 \qquad Ex^2 = 0$$

Thus we must have

$$A = 0 \qquad\qquad B = D \qquad\qquad E = 0$$

With these values of A, B, C, and D, eqq. (10.45) reduce to

$$c\,dt = B\,d\alpha\,x \qquad\qquad dx = B\,d\alpha\,ct \tag{10.47}$$

If now we define a new variable $\tilde{\alpha}$ by

$$d\tilde{\alpha} = B\,d\alpha$$

then eqq. (10.47) further simplify to

$$c\,dt = d\tilde{\alpha}\,x \qquad\qquad dx = d\tilde{\alpha}\,ct$$

or

$$\frac{d(ct)}{d\tilde{\alpha}} = x \qquad\qquad \frac{dx}{d\tilde{\alpha}} = ct \tag{10.48}$$

Eqq. (10.48) constitute a set of coupled differential equations for x and t. We will now solve those equations by two methods, one that uses matrices (for those of you familiar with matrices) and one that doesn't (for those of you unfamiliar with them).

Method 1: With Matrices

If we define the column vector $\mathbf{x}$ by

$$\mathbf{x} = \begin{pmatrix} ct \\ x \end{pmatrix}$$

then eqq. (10.48) can be written

$$\frac{d\mathbf{x}}{d\tilde{\alpha}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \mathbf{x} \tag{10.49}$$

As you can verify by separating variables and integrating, the solution to an equation of the form

$$\frac{df}{dz} = kf \qquad (k = \text{const})$$

is

$$f = e^{kz} f_0$$

where f_0 is the "initial" value of f, that is, the value of f at $z = 0$. The solution in the case of a matrix equation is no different: eq. (10.49) will yield simply

$$\mathbf{x} = e^{\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\tilde{\alpha}} \mathbf{x}_0 \tag{10.50}$$

where $\mathbf{x}_0$ is the value of $\mathbf{x}$ at $\tilde{\alpha} = 0$. Recalling that $d\tilde{\alpha}$ generated the change in coordinates as we went from frame F to frame F', eq. (10.50) is telling us that the exponential of the above matrix takes us from the coordinates (x, t) of F to the coordinates (x', t') of F'.[34] That is, if

$$\mathbf{x}' = \begin{pmatrix} ct' \\ x' \end{pmatrix} \qquad \text{and} \qquad \mathbf{x} = \begin{pmatrix} ct \\ x \end{pmatrix}$$

then

$$\mathbf{x}' = e^{\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\tilde{\alpha}}\mathbf{x} \tag{10.51}$$

We can do out the exponential of the matrix in eq. (10.51) by noting that it has a special property:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Thus any even power of this matrix will equal the identity matrix, and any odd power, being the product of the matrix with some even power of itself, will equal the matrix itself:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^{2n} = \left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^2\right)^n = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^n = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^{2n+1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^{2n}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

So for the Taylor expansion of the exponential we have

$$\begin{aligned} e^{\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)\tilde{\alpha}} &= \sum_{n=0}^{\infty} \frac{\tilde{\alpha}^n}{n!}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^n \\ &= \sum_{n \text{ even}} \frac{\tilde{\alpha}^n}{n!}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^n + \sum_{n \text{ odd}} \frac{\tilde{\alpha}^n}{n!}\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^n \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\sum_{n \text{ even}} \frac{\tilde{\alpha}^n}{n!} + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\sum_{n \text{ odd}} \frac{\tilde{\alpha}^n}{n!} \end{aligned}$$

To see how the sums that we are left with work out, consider the Taylor expansions of the definitions of cosh and sinh:

$$\cosh\tilde{\alpha} = \tfrac{1}{2}(e^{\tilde{\alpha}} + e^{-\tilde{\alpha}}) \qquad\qquad \sinh\tilde{\alpha} = \tfrac{1}{2}(e^{\tilde{\alpha}} - e^{-\tilde{\alpha}})$$

[34] The matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ is known as the generator of the Lorentz transform, in the sense that it generates an infinitesimal Lorentz transform. For any kind of continuous transform, if you know the generators of infinitesimal transforms, you can integrate over a succession of such infinitesimal transforms to obtain a result for finite transforms. These integrations yield exponentials of the generators, similar to our result (10.50).

For the cosh, the terms in the expansions of the positive and negative exponentials will cancel for odd powers of $\tilde{\alpha}$, leaving just double the sum over the even powers — a factor of two neatly canceled by the $\frac{1}{2}$ in the definition of the cosh. For the sinh, the same cancellation will occur for the even powers. Thus

$$\cosh\tilde{\alpha} = \sum_{n \text{ even}} \frac{\tilde{\alpha}^n}{n!} \qquad\qquad \sinh\tilde{\alpha} = \sum_{n \text{ odd}} \frac{\tilde{\alpha}^n}{n!}$$

and our expansion of the exponential of the matrix reduces to

$$e^{\left(\begin{smallmatrix}0&1\\1&0\end{smallmatrix}\right)\tilde{\alpha}} = \begin{pmatrix}1&0\\0&1\end{pmatrix}\cosh\tilde{\alpha} + \begin{pmatrix}0&1\\1&0\end{pmatrix}\sinh\tilde{\alpha} = \begin{pmatrix}\cosh\tilde{\alpha}&\sinh\tilde{\alpha}\\\sinh\tilde{\alpha}&\cosh\tilde{\alpha}\end{pmatrix}$$

Pulling out an overall factor of $\cosh\tilde{\alpha}$ and noting that by definition

$$\tanh\tilde{\alpha} = \frac{\sinh\tilde{\alpha}}{\cosh\tilde{\alpha}}$$

we find that eq. (10.51) consequently simplifies to

$$\mathbf{x}' = \begin{pmatrix}\cosh\tilde{\alpha}&\sinh\tilde{\alpha}\\\sinh\tilde{\alpha}&\cosh\tilde{\alpha}\end{pmatrix}\mathbf{x} = \cosh\tilde{\alpha}\begin{pmatrix}1&\tanh\tilde{\alpha}\\\tanh\tilde{\alpha}&1\end{pmatrix}\mathbf{x} \tag{10.52a}$$

Expressing this relation explicitly component by component, we have

$$\begin{pmatrix}ct'\\x'\end{pmatrix} = \cosh\tilde{\alpha}\begin{pmatrix}1&\tanh\tilde{\alpha}\\\tanh\tilde{\alpha}&1\end{pmatrix}\begin{pmatrix}ct\\x\end{pmatrix} \tag{10.52b}$$

or

$$\begin{aligned} ct' &= \cosh\tilde{\alpha}\,(ct + \tanh\tilde{\alpha}\,x) \\ x' &= \cosh\tilde{\alpha}\,(\tanh\tilde{\alpha}\,ct + x) \end{aligned} \tag{10.52c}$$

Eq. (10.52a), or, equivalently, eqq. (10.52c), is the most general mathematical result for the symmetry operation that leaves $c^2t^2 - x^2$ invariant; it is the most general possible form of the Lorentz transform.[35]

If you followed this, you can now skip Method 2 and go to the section marked "Back to Business" on p.547. If you didn't follow this, you will probably find Method 2 easier to digest. Maybe. We hope.

Method 2: Without Matrices

Our object is to solve eqq. (10.48):

$$\frac{d(ct)}{d\tilde{\alpha}} = x \qquad\qquad \frac{dx}{d\tilde{\alpha}} = ct \tag{10.48}$$

[35] Modulo the considerations noted in footnote 33 on p.540.

In order to integrate these equations for x and t, we need to separate these coupled differential equations that mix x and t into an equation involving only x and another equation involving only t. The trick to accomplishing this is to note that the latter of eqq. (10.48) gives $dx/d\tilde{\alpha}$ in terms of t, and so if we take a derivative of the former of eqq. (10.48) we will be able to substitute and obtain a relation in terms solely of t:

$$\frac{d^2(ct)}{d\tilde{\alpha}^2} = \frac{dx}{d\tilde{\alpha}} = ct \tag{10.53}$$

Using the very same trick for the latter of eqq. (10.48), we also have

$$\frac{d^2x}{d\tilde{\alpha}^2} = \frac{d(ct)}{d\tilde{\alpha}} = x \tag{10.54}$$

Eqq. (10.53) and (10.54) are of the form

$$\frac{d^2f}{dz^2} = k^2 f \qquad (k = \text{const})$$

The solution to such equations, as you can verify by substitution, is

$$f = Pe^{kz} + Qe^{-kz}$$

where P and Q are arbitrary constants. In eqq. (10.53) and (10.54), the constant k is simply 1, so that the solutions are

$$\begin{aligned} ct &= P_t e^{\tilde{\alpha}} + Q_t e^{-\tilde{\alpha}} \\ x &= P_x e^{\tilde{\alpha}} + Q_x e^{-\tilde{\alpha}} \end{aligned} \tag{10.55}$$

where P_t, Q_t, P_x, and Q_x are constants, the values of which we must now determine.

First of all, eqq. (10.55) must satisfy the original coupled equations (10.48). Substituting the solutions (10.55) into eqq. (10.48), we have

$$\begin{aligned} \frac{d(ct)}{d\tilde{\alpha}} &= x \\ \frac{d}{d\tilde{\alpha}}(P_t e^{\tilde{\alpha}} + Q_t e^{-\tilde{\alpha}}) &= P_x e^{\tilde{\alpha}} + Q_x e^{-\tilde{\alpha}} \\ P_t e^{\tilde{\alpha}} - Q_t e^{-\tilde{\alpha}} &= P_x e^{\tilde{\alpha}} + Q_x e^{-\tilde{\alpha}} \end{aligned}$$

and

$$\begin{aligned} \frac{dx}{d\tilde{\alpha}} &= ct \\ \frac{d}{d\tilde{\alpha}}(P_x e^{\tilde{\alpha}} + Q_x e^{-\tilde{\alpha}}) &= P_t e^{\tilde{\alpha}} + Q_t e^{-\tilde{\alpha}} \\ P_x e^{\tilde{\alpha}} - Q_x e^{-\tilde{\alpha}} &= P_t e^{\tilde{\alpha}} + Q_t e^{-\tilde{\alpha}} \end{aligned}$$

Since these equations must hold for all values of $\tilde{\alpha}$, and since $e^{\tilde{\alpha}}$ and $e^{-\tilde{\alpha}}$ are functionally independent, the equality must hold separately for the terms in $e^{\tilde{\alpha}}$ and the terms in $e^{-\tilde{\alpha}}$:

$$\begin{aligned} P_t &= P_x \qquad \text{and} \qquad -Q_t = Q_x \\ P_x &= P_t \qquad \text{and} \qquad -Q_x = Q_t \end{aligned}$$

Thus we have two constraints, $P_x = P_t$ and $Q_x = -Q_t$. Applying these constraints to eqq. (10.55), we have

$$\begin{aligned} ct &= P_t e^{\tilde{\alpha}} + Q_t e^{-\tilde{\alpha}} \\ x &= P_t e^{\tilde{\alpha}} - Q_t e^{-\tilde{\alpha}} \end{aligned} \tag{10.56}$$

To make further progress, we must be more precise about what eqq. (10.56) mean. If we remember that $d\tilde{\alpha}$ generated the change in coordinates as we went from frame F to frame F', the right-hand sides of eqq. (10.56) are telling us the coordinates that result from the Lorentz transform, that is, the primed coordinates of frame F':

$$\begin{aligned} ct' &= P_t e^{\tilde{\alpha}} + Q_t e^{-\tilde{\alpha}} \\ x' &= P_t e^{\tilde{\alpha}} - Q_t e^{-\tilde{\alpha}} \end{aligned} \tag{10.57}$$

We therefore need to relate P_t and Q_t to the coordinates x and t of frame F. Recall now that when β is small, so is α. In particular, $\alpha = 0$ when the relative velocity between the frames F and F' vanishes, and in this case, the two frames being the same frame, the coordinates of F and F' are the same: $t' = t$ and $x' = x$. Setting $t' = t$, $x' = x$, and $\tilde{\alpha} = 0$ in eqq. (10.57) gives

$$\begin{aligned} ct &= P_t + Q_t \\ x &= P_t - Q_t \end{aligned}$$

We thus have two equations in the two unknowns P_t and Q_t. The sum and difference of these two equations yield, respectively,

$$\begin{aligned} ct + x &= 2P_t \\ ct - x &= 2Q_t \end{aligned}$$

so that $P_t = \frac{1}{2}(ct + x)$ and $Q_t = \frac{1}{2}(ct - x)$. Using these results for P_t and Q_t in eqq. (10.57) we arrive at

$$\begin{aligned} ct' &= \tfrac{1}{2}(ct + x)e^{\tilde{\alpha}} + \tfrac{1}{2}(ct - x)e^{-\tilde{\alpha}} \\ x' &= \tfrac{1}{2}(ct + x)e^{\tilde{\alpha}} - \tfrac{1}{2}(ct - x)e^{-\tilde{\alpha}} \end{aligned}$$

or, grouping together terms in ct and terms in x,

$$\begin{aligned} ct' &= \tfrac{1}{2}(e^{\tilde{\alpha}} + e^{-\tilde{\alpha}})\, ct + \tfrac{1}{2}(e^{\tilde{\alpha}} - e^{-\tilde{\alpha}})\, x \\ x' &= \tfrac{1}{2}(e^{\tilde{\alpha}} - e^{-\tilde{\alpha}})\, ct + \tfrac{1}{2}(e^{\tilde{\alpha}} + e^{-\tilde{\alpha}})\, x \end{aligned} \tag{10.58}$$

Recalling the definitions of the hyperbolic functions

$$\cosh\tilde{\alpha} = \tfrac{1}{2}(e^{\tilde{\alpha}} + e^{-\tilde{\alpha}}) \qquad\qquad \sinh\tilde{\alpha} = \tfrac{1}{2}(e^{\tilde{\alpha}} - e^{-\tilde{\alpha}})$$

we can more neatly write eqq. (10.58) as

$$\begin{aligned} ct' &= \cosh\tilde{\alpha}\; ct + \sinh\tilde{\alpha}\; x \\ x' &= \sinh\tilde{\alpha}\; ct + \cosh\tilde{\alpha}\; x \end{aligned}$$

If we pull out an overall factor of $\cosh\tilde{\alpha}$ and note that by definition

$$\tanh\tilde{\alpha} = \frac{\sinh\tilde{\alpha}}{\cosh\tilde{\alpha}}$$

we finally arrive at

$$\begin{aligned} ct' &= \cosh\tilde{\alpha}\; (ct + \tanh\tilde{\alpha}\; x) \\ x' &= \cosh\tilde{\alpha}\; (\tanh\tilde{\alpha}\; ct + x) \end{aligned} \tag{10.52c}$$

Eqq. (10.52c) are the most general mathematical result for the symmetry operation that leaves $c^2t^2 - x^2$ invariant; it is the most general possible form of the Lorentz transform.[36]

Back to Business

Our remaining task is to relate the still unknown parameter $\tilde{\alpha}$ to the physical parameter $\beta = v/c$ that specifies the relative motion of the frames F and F'. To accomplish this, we need merely work out the Lorentz transform in terms of β for a trivially simple case and match this result to eqq. (10.52c). It will suffice to consider an object at rest at $x = 0$ in frame F. If F' is moving at velocity v relative to F, then in frame F' the object will appear to be moving backward, with velocity $-v$. Since we have set up our frames F and F' to have a common spacetime origin, so that $x = 0$, $t = 0$ coincides with $x' = 0$, $t' = 0$, the object, which in frame F is always (and therefore in particular at time $t = 0$) at $x = 0$, will be at $x' = 0$ at time $t' = 0$. At time t' (corresponding to time t in frame F), the object will therefore be at $x' = -vt'$ in frame F'. Writing

$$x' = -vt' = -\frac{v}{c}\, ct' = -\beta\, ct'$$

and using the points $(x, t) = (0, t)$ and $(x', t') = (-\beta\, ct',\ t')$ in eqq. (10.52c), we have

$$\begin{aligned} ct' &= \cosh\tilde{\alpha}\; ct \\ -\beta\, ct' &= \cosh\tilde{\alpha}\;\; \tanh\tilde{\alpha}\; ct \end{aligned}$$

[36] Modulo the considerations noted in footnote 33 on p.540.

Dividing the latter relation by the former yields

$$\tanh\tilde{\alpha} = -\beta \tag{10.59}$$

Now, as you can easily verify by explicit substitution of the definitions

$$\cosh\tilde{\alpha} = \tfrac{1}{2}(e^{\tilde{\alpha}} + e^{-\tilde{\alpha}}) \qquad\qquad \sinh\tilde{\alpha} = \tfrac{1}{2}(e^{\tilde{\alpha}} - e^{-\tilde{\alpha}})$$

the hyperbolic equivalent of

$$\sin^2\theta + \cos^2\theta = 1$$

is

$$\cosh^2\tilde{\alpha} - \sinh^2\tilde{\alpha} = 1$$

Dividing this hyperbolic relation by $\cosh^2\tilde{\alpha}$ gives

$$1 - \tanh^2\tilde{\alpha} = \frac{1}{\cosh^2\tilde{\alpha}}$$

Solving this relation for $\cosh\tilde{\alpha}$ and using $\tanh\tilde{\alpha} = -\beta$, we arrive at

$$\cosh\tilde{\alpha} = \frac{1}{\sqrt{1-\tanh^2\tilde{\alpha}}} = \frac{1}{\sqrt{1-\beta^2}} = \gamma \tag{10.60}$$

Eqq. (10.52c), expressed in terms of β, are thus the same as eqq. (10.29):

$$\begin{aligned} ct' &= \gamma(ct - \beta x) \\ x' &= \gamma(-\beta\, ct + x) \end{aligned}$$

Voilà.

Although we should, for completeness, say something about the Lorentz transform of the transverse (that is, the y and z) directions. To see what happens along these directions, it is easiest to work in polar coordinates, with the radial coordinate r being measured from our x axis. Now, we expect that space is symmetric, that is, that there is no special or preferred direction. Although our setting things up so that the relative motion of the two frames F and F' was along the x direction broke this symmetry and made the x direction special, the symmetry of the y and z directions is still unbroken: there is no reason for there to be any preferred direction in the yz plane. A dependence on the polar angle θ is therefore ruled out. And a dependence on the radial direction is similarly ruled out: r must be measured from an axis, but the only constraint is that this axis run parallel to our x axis: the relative motion of F and F' distinguishes only an x *direction* and not a specific choice of x *axis*. Since any axis parallel to the x axis must be as valid as any other, and since the radial coordinates of such axes will differ,

there can be no dependence on the radial coordinate. The Lorentz transform therefore affects only the time coordinate and the spatial coordinate *along the direction of motion* — more precisely, along the direction of the relative motion of the two frames. For relative motion along the x direction, the full result for the Lorentz transform is thus

$$\begin{aligned} ct' &= \gamma(ct - \beta x) \\ x' &= \gamma(-\beta\, ct + x) \\ y' &= y \\ z' &= z \end{aligned} \tag{10.61}$$

or, in terms of matrices,

$$\begin{pmatrix} ct' \\ x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \gamma & -\beta\gamma & 0 & 0 \\ -\beta\gamma & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \\ y \\ z \end{pmatrix} \tag{10.62}$$

10.3.8 Lorentz Transforms as Rotations

Lorentz transforms can be regarded as rotations that involve a space and the time direction instead of two spatial directions. To see how this is so, we can apply the techniques of the preceding section to discover the continuous transforms (other than translations) that will leave distances in the xy plane invariant. The squared distance s^2 between the origin and a point (x, y) is given by the Pythagorean relation

$$s^2 = x^2 + y^2$$

Following the reasoning of the preceding section, we consider an infinitesimal transform on the coordinates x and y. Such a transform must shift the values of x and y by some infinitesimal linear combination of x and y:

$$\begin{aligned} dx &= A\, d\alpha\, x + B\, d\alpha\, y \\ dy &= D\, d\alpha\, x + E\, d\alpha\, y \end{aligned} \tag{10.63}$$

where A, B, C, and D are as yet unknown coefficients and α is some parameter that is small for an infinitesimal transform. If we now impose the condition that the distance s not change (that is, that $ds = 0$), we have $d(s^2) = 2s\, ds = 0$ and hence

$$\begin{aligned} d(s^2) &= d(x^2) + d(y^2) \\ 0 &= 2x\, dx + 2y\, dy \end{aligned}$$

Substituting eqq. (10.63) into this relation and recombining terms in x^2, xy, and y^2, we have

$$\begin{aligned} 0 &= 2x(A\,d\alpha\,x + B\,d\alpha\,y) + 2y(D\,d\alpha\,x + E\,d\alpha\,y) \\ &= 2\,d\alpha\Big(Ax^2 + (B+D)xy + Ey^2\Big) \end{aligned}$$

Since this relation must hold for all x and y, the equality must separately hold for the terms in x^2, the terms in xy, and the terms in y^2. We therefore have

$$A = 0 \qquad\qquad B + D = 0 \qquad\qquad E = 0$$

Eqq. (10.63) are thus reduced to

$$\begin{aligned} dx &= B\,d\alpha\,y \\ dy &= -B\,d\alpha\,x \end{aligned}$$

As before, we absorb the coefficient B into the α by defining

$$d\tilde{\alpha} = B\,d\alpha$$

so that eqq. (10.63) further reduce to

$$\begin{aligned} dx &= d\tilde{\alpha}\,y \\ dy &= -\,d\tilde{\alpha}\,x \end{aligned} \tag{10.64}$$

or, in matrix form,

$$\begin{pmatrix} dx \\ dy \end{pmatrix} = d\tilde{\alpha} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{10.65}$$

Dividing by $d\tilde{\alpha}$ on both sides, we arrive at

$$\frac{d\mathbf{x}}{d\tilde{\alpha}} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \mathbf{x} \tag{10.66}$$

where we have defined the column vector

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}$$

As before, the solution to this equation is

$$\mathbf{x} = e^{\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)\tilde{\alpha}} \mathbf{x}_0 \tag{10.67}$$

where $\mathbf{x}_0$ is the value of $\mathbf{x}$ at $\tilde{\alpha} = 0$. In other words, the transform that takes us from our original coordinates x and y to the new coordinates x' and y' is the exponential of the matrix, so that if

$$\mathbf{x}' = \begin{pmatrix} x' \\ y' \end{pmatrix} \qquad \text{and} \qquad \mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}$$

then

$$\mathbf{x}' = e^{\left(\begin{smallmatrix}0 & 1\\ -1 & 0\end{smallmatrix}\right)\tilde{\alpha}}\mathbf{x} \tag{10.68}$$

We can do out the exponential of the matrix in eq. (10.68) by noting that it has the property

$$\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}^2 = \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix} = -\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

The even and odd powers of this matrix will thus work out to

$$\begin{aligned}
\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}^{2n} &= \left(\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}^2\right)^n \\
&= \left(-\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)^n \\
&= (-1)^n\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} \\
\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}^{2n+1} &= \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}^{2n}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix} \\
&= (-1)^n\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix} \\
&= (-1)^n\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}
\end{aligned}$$

So for the Taylor expansion of the exponential we have

$$\begin{aligned}
e^{\left(\begin{smallmatrix}0 & 1\\ -1 & 0\end{smallmatrix}\right)\tilde{\alpha}} &= \sum_{n=0}^{\infty}\frac{1}{n!}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}^n \\
&= \sum_{n\ \text{even}}\frac{\tilde{\alpha}^n}{n!}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}^n + \sum_{n\ \text{odd}}\frac{\tilde{\alpha}^n}{n!}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}^n \\
&= \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\sum_{n\ \text{even}}\frac{(-1)^{\frac{n}{2}}\tilde{\alpha}^n}{n!} + \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\sum_{n\ \text{odd}}\frac{(-1)^{\frac{n-1}{2}}\tilde{\alpha}^n}{n!}
\end{aligned}$$

The first sum on the right-hand side is over even powers $\tilde{\alpha}^n/n!$ with alternating sign — that is, the series expansion of the cosine. The second sum on the right-hand side is over odd powers $\tilde{\alpha}^n/n!$ with alternating sign — the series expansion of the sine. We therefore have

$$e^{\left(\begin{smallmatrix}0 & 1\\ -1 & 0\end{smallmatrix}\right)\tilde{\alpha}} = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\cos\tilde{\alpha} + \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\sin\tilde{\alpha} = \begin{pmatrix}\cos\tilde{\alpha} & \sin\tilde{\alpha}\\ -\sin\tilde{\alpha} & \cos\tilde{\alpha}\end{pmatrix}$$

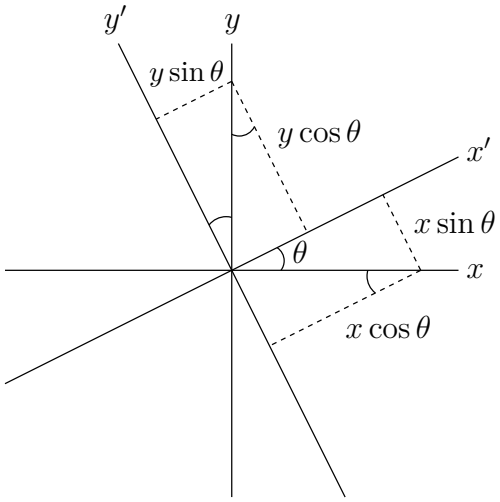


Figure 10.6: Rotated Axes

so that eq. (10.68) simplifies to

$$\mathbf{x}' = \begin{pmatrix} \cos\tilde{\alpha} & \sin\tilde{\alpha} \\ -\sin\tilde{\alpha} & \cos\tilde{\alpha} \end{pmatrix} \mathbf{x} \tag{10.69}$$

And there you have it: this is our result for the transform that leaves distances in the xy plane invariant.

To relate transforms of the form (10.69) to rotations, consider the rotation of the coordinate axes shown in fig. (10.6). As you can see from the dashed lines,

$$\begin{aligned} x' &= \cos\theta\, x + \sin\theta\, y \\ y' &= -\sin\theta\, x + \cos\theta\, y \end{aligned} \tag{10.70}$$

Eqq. (10.70) are identical with eq. (10.69) if we identify $\tilde{\alpha}$ with the angle θ of rotation.

To see how Lorentz transforms can be regarded as rotations, recall that Lorentz transforms leave

$$c^2t^2 - x^2$$

invariant, while the quantity left invariant by rotations is

$$x^2 + y^2$$

The former involves the difference of two squares while the latter involves the sum. The difference $c^2t^2 - x^2$ can, however, be made to look like the sum $x^2 + y^2$ if we make a simple change of variables: if we define

$$\tilde{t} = it$$

where $i = \sqrt{-1}$, then $t = -i\tilde{t}$ and

$$c^2t^2 - x^2 = c^2(-i\tilde{t}\,)^2 - x^2 = -\left(c^2\tilde{t}^{\,2} + x^2\right)$$

Now, the transforms that leave $-\left(c^2\tilde{t}^{\,2} + x^2\right)$ invariant are of course the same as those that leave $c^2\tilde{t}^{\,2} + x^2$ invariant, so that Lorentz transforms must be a sort of rotation:

$$\begin{pmatrix} c\tilde{t}\,' \\ x' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} c\tilde{t} \\ x \end{pmatrix} \tag{10.71}$$

Our remaining task is to figure out what the angle θ of the rotation between the spatial axis x and the imaginary-time axis $\tilde{t}$ corresponds to physically. First, we need to revert from imaginary time $\tilde{t}$ to the physically real time t: substituting it for $\tilde{t}$ in eq. (10.71), we have

$$\begin{pmatrix} ict\,' \\ x' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} ict \\ x \end{pmatrix} \tag{10.72}$$

Next, to compare this to the Lorentz transform

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \gamma \begin{pmatrix} 1 & -\beta \\ -\beta & 1 \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix} \tag{10.73}$$

we want to rewrite eq. (10.72) in terms of the column vectors

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} ct \\ x \end{pmatrix}$$

If we look at the result of the matrix multiplication in eq. (10.72),

$$\begin{aligned} ict' &= \cos\theta\; ict + \sin\theta\; x \\ x' &= -\sin\theta\; ict + \cos\theta\; x \end{aligned}$$

we see that if we divide the top equation by i, we will be left with the desired quantities (ct' and x') on the left-hand side:

$$\begin{aligned} ct' &= \cos\theta\; ct - i\sin\theta\; x \\ x' &= -\sin\theta\; ict + \cos\theta\; x \end{aligned}$$

We can then rewrite these relations in matrix form in terms of the column vectors (ct', x') and (ct, x) as

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \begin{pmatrix} \cos\theta & -i\sin\theta \\ -i\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} ct \\ x \end{pmatrix} \tag{10.74}$$

We could now compare eq. (10.74) and Lorentz transform (10.73), but the relation between the angle θ of rotation and β will, as it turns out, be more

transparent if we further rewrite eq. (10.74) in terms of hyperbolic functions and compare it to our results (10.52b), (10.59), and (10.60) (see pp.544 and 548) for the Lorentz transform:

$$\begin{pmatrix} ct' \\ x' \end{pmatrix} = \cosh\tilde{\alpha}\begin{pmatrix} 1 & \tanh\tilde{\alpha} \\ \tanh\tilde{\alpha} & 1 \end{pmatrix}\begin{pmatrix} ct \\ x \end{pmatrix} \qquad \text{with} \quad \begin{matrix} \beta = -\tanh\alpha \\ \gamma = \cosh\alpha \end{matrix} \tag{10.75}$$

Consider the hyperbolic functions of an imaginary angle $i\varphi$:

$$\begin{aligned} \cosh i\varphi &= \tfrac{1}{2}(e^{i\varphi} + e^{-i\varphi}) \\ &= \tfrac{1}{2}\Big((\cos\varphi + i\sin\varphi) + (\cos\varphi - i\sin\varphi)\Big) \\ &= \cos\varphi \\ \sinh i\varphi &= \tfrac{1}{2}(e^{i\varphi} - e^{-i\varphi}) \\ &= \tfrac{1}{2}\Big((\cos\varphi + i\sin\varphi) - (\cos\varphi - i\sin\varphi)\Big) \\ &= i\sin\varphi \end{aligned}$$

We can therefore write

$$\cos\theta = \cosh i\theta \qquad\qquad i\sin\theta = \sinh i\theta$$

in terms of which the transform of eq. (10.74) becomes

$$\begin{pmatrix} \cosh i\theta & -\sinh i\theta \\ -\sinh i\theta & \cosh i\theta \end{pmatrix} = \cosh i\theta\begin{pmatrix} 1 & -\tanh i\theta \\ -\tanh i\theta & 1 \end{pmatrix} \tag{10.76}$$

Comparing eq. (10.76) with eqq. (10.75), we see that $\tilde{\alpha} = -i\theta$ and $\beta = \tanh i\theta$. Thus we arrive at

$$\theta = i\tilde{\alpha} = -i\tanh^{-1}\beta$$

The Lorentz transform can therefore be regarded as a rotation by an imaginary angle.

In fact, the fundamental difference between time and space is simply the negative sign that occurs in $c^2t^2 - x^2$. If one works in terms of an imaginary time coordinate, this difference becomes a sum and the geometry of spacetime becomes Euclidean: Lorentz transforms become simply a rotations between time and space, in no way different from rotations between two spatial axes. With a real time coordinate, spacetime is non-Euclidean, but Lorentz transforms can still be regarded as rotations between time and space — it is just that the angle of rotation is imaginary. The full set of six rotations in four-dimensional spacetime (the spatial rotations about the three spatial axes, plus the rotations between the time axis and each of the three spatial axes) constitutes what is known as the Lorentz group. These, together with

the four translations or shifts along the four spacetime axes, constitute the Poincaré group. For a physical theory to be valid, it must exhibit Lorentz (and, more generally, Poincaré) invariance. That is, the laws postulated by the theory must hold for all frames, regardless of their location or orientation in spacetime. It is in fact the condition of Lorentz invariance that constrains string theory to ten spacetime dimensions; in other numbers of dimensions, the theory would violate Lorentz invariance.[37]

10.4 Time Dilation & Length Contraction

When deriving basic results in relativity, it is very important to think in terms of events rather than just distances and times. In the Lorentz transform, we see that both Δx and Δt contribute to $\Delta x'$ and also to $\Delta t'$. That is, what is a spatial distance to one observer is partly a time interval to another and vice versa. While in Newtonian physics time is absolute, in relativity space and time are coupled inextricably together as four-dimensional spacetime and it makes no sense to consider a spatial displacement by itself or a time interval by itself.

10.4.1 Time Dilation

Consider a clock at rest in the moving (primed) frame. To see how the time elapsed in this moving frame compares to the time elapsed in the our stationary (unprimed) frame, we will take as the two events two ticks of the clock. Since in the moving frame the clock is at rest, we have $\Delta x' = 0$, with $\Delta t'$ being the time between ticks according to the clock. In our frame, the time between ticks is Δt, and since to us the clock is traveling at velocity v between ticks, the displacement between ticks is $\Delta x = v\Delta t$. Using the Lorentz transform relation (10.29b) and recalling that $\beta = v/c$, we thus have

$$\begin{aligned} c\Delta t' &= \gamma(-\beta\Delta x + c\Delta t) \\ &= \gamma(-\beta v\Delta t + c\Delta t) \\ &= \gamma\left(-\beta\frac{v}{c} + 1\right)\, c\Delta t \\ &= \gamma(-\beta^2 + 1)\, c\Delta t \\ &= \gamma(1-\beta^2)\, c\Delta t \end{aligned}$$

[37] This, along with the fact that the sum of the positive integers equals $-\frac{1}{12}$:

$$\sum_{n=1}^{\infty} n = 1 + 2 + 3 + 4 + \cdots = -\tfrac{1}{12}$$

For those of you who are incredulous, the proof is given in Appendix E.

Now, from eq. (10.28),

$$\gamma = \frac{1}{\sqrt{1-\beta^2}}$$

which, if we invert and square both sides, gives

$$\frac{1}{\gamma^2} = 1 - \beta^2 \tag{10.77}$$

Thus our relation for $\Delta t'$ can be written

$$c\Delta t' = \gamma \, \frac{1}{\gamma^2} \, c\Delta t$$

or

$$\Delta t' = \frac{1}{\gamma} \, \Delta t$$

Since $\gamma \geq 1$, $\Delta t' \leq \Delta t$: the time $\Delta t'$ between ticks according to the clock is less than the time Δt between ticks according to us. Stated more simply, less time elapses in the moving frame of the clock. This effect is known as time dilation:

> **Time Dilation**: A moving clock ticks more slowly than a clock at rest by a factor of γ.

Note that this effect is a property of spacetime itself, not of the clock: it is time itself that slows down, so that the effect is the same for mechanical, electric, biological, and all other kinds of clocks.

If we had the means to accelerate rocket ships to speeds close to the speed of light — a feat far exceeding our current capabilities and, as will be explained when we get to relativistic results for energy, likely to remain forever beyond them — , time dilation would make travel to distant stars humanly possible. The stars we'd want to visit are likely to be hefty number of light-years away from us.[38] Now, a light-year is the distance light travels in one year, so a star 1000 light-years away would take, by definition, 1000 years to reach traveling at the speed of light. As we will see when we get to relativistic results for energy, no object traveling at less than the speed of light can ever be accelerated all the way up to the speed of light — if you put enough energy into it, you can get it as close as you want to the speed of light, but you can't ever actually reach the speed of light. And traveling at close to the speed of light, it would take a rocket ship slightly more than 1000

[38] The star closest to us (excepting the Sun, for the wisenheimers among you) is Alpha Centauri, which is about 4.3 light-years away, and there are quite a number of stars in the 5–20 light-year range, but these are close enough that we can see they wouldn't really be all that interesting to visit. To find a star with an interesting planetary system we'd have to go way out there.

years to reach a star 1000 light-years away. Not very practical, you would think — 40 generations would pass on the way there, not to mention the way back. But if we could boost the rocket ship to a speed so close to that of light that its γ were, say, 1000, because of time dilation only a little over a year would elapse on the ship during the journey. At speeds even closer to that of light, γ would be even larger and the trip even shorter. Instead of taking 80 generations, a round trip could be made in a little over two years — even less with a larger γ.

This would, however, not be a great boon to science, because time is dilated only in the moving frame, that is, for the ship and for the people and things on the ship; here on Earth, time would be elapsing at its "normal" rate, with the result that the round trip to the star and back would take a little over 2000 years by our clocks. The returning astronauts would step off the ship having aged a little more than two years during the round trip, but 80 generations would have passed on Earth. While it would be interesting to meet your great great great

step off the ship and, like in a bad movie, suddenly turn into mummies and cobwebbed skeletons and then disintegrate into dust that blows away in the wind. Although this difference in elapsed time is, like other relativistic effects, imperceptibly small at everyday velocities, it is always present: every time you run a lap around the track, you have aged less than your coach.

10.4.2 Length Contraction

Our derivation of the time dilation effect was an example of what is called a thought (or, from the German, gedanken) experiment: we perform an imaginary experiment and reason out the result we should get according to relativistic principles.

Now consider another thought experiment: we determine the length of a moving meter stick. In relativity, we must prescribe the method for determining the length of the stick in terms of events: it is important not just to look at where the ends of the stick are located, but to look at the two ends *at the same time*. In the moving frame, where the stick is at rest, this makes no difference, but it does make a difference in our frame, where the stick appears to be moving at velocity v: if we aren't careful to mark both ends of the stick at the same time, our measurement of its length will include the distance it has moved in the meantime. We will therefore take as our two events

1. The marking of the back end of the stick, and

2. The marking of the front end of the stick.

These two markings must be done simultaneously if the distance between marks is to accurately reflect the length of the stick itself.

In the moving frame, marking the two ends simultaneously means that there is no time difference $\Delta t'$ between markings: $\Delta t' = 0$. The distance between markings is then the actual length of the stick, which we will denote by ℓ':

$$\Delta x' = x'_2 - x'_1 = x'_{front} - x'_{back} = \ell'$$

Since the stick is at rest in this frame, this length ℓ' is called the rest length of the stick. Using $\Delta t' = 0$ and $\Delta x' = \ell'$ in the Lorentz transform (10.29), we have [39]

$$\ell' = \gamma(\Delta x - \beta c \Delta t) \tag{10.78a}$$
$$0 = \gamma(-\beta \Delta x + c \Delta t) \tag{10.78b}$$

[39] We could get results for Δt and Δx directly, without the need for any algebra, from the inverse Lorentz transform (10.36), but, hey, you only go around once.

If we formally solve the latter of these two relations for the time Δt between markings in our frame, we obtain

$$\Delta t = \frac{1}{c}\beta\Delta x$$

which is clearly *not* zero. This is special case of a general result:

> **Simultaneity**: Events that are simultaneous in one reference frame are in general *not* simultaneous in another reference frame.

Using this result $\Delta t = \frac{1}{c}\beta\Delta x$ in eq. (10.78a), we obtain

$$\begin{aligned}\ell' &= \gamma(\Delta x - \beta c\Delta t)\\ &= \gamma\Big(\Delta x - \beta c\cdot\frac{1}{c}\beta\Delta x\Big)\\ &= \gamma(\Delta x - \beta^2\Delta x)\\ &= \gamma(1-\beta^2)\Delta x\end{aligned}$$

which, if we use $1-\beta^2 = 1/\gamma^2$ from eq. (10.77), becomes

$$\ell' = \gamma\frac{1}{\gamma^2}\Delta x$$

or

$$\Delta x = \gamma\ell' \tag{10.79}$$

Since $\gamma \geq 1$, it would seem that we find the stick to be longer, but we must be careful: we have already found that the two ends of the stick are not marked simultaneously in our frame. To obtain the length of the stick from our result for Δx, we must therefore take into account that, after we marked the back of the stick, the front of the stick was moving forward at velocity v for a time

$$\Delta t = t_2 - t_1 = t_{front} - t_{back} = \frac{1}{c}\beta\Delta x$$

before it was marked, as shown in fig. (10.7). During this time, distance

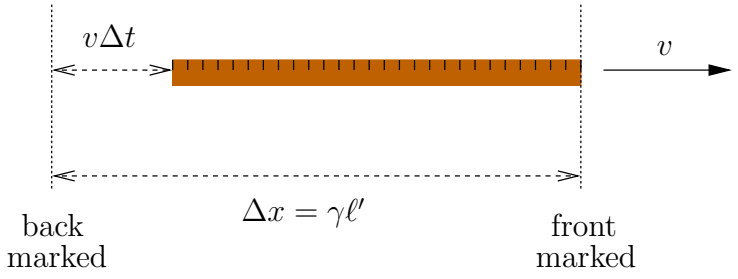


Figure 10.7: A Shifty Stick

that the front of the stick moved forward was

$$v\Delta t = v\left(\frac{1}{c}\beta\Delta x\right) = \frac{v}{c}\beta\Delta x = \beta^2\Delta x$$

We need to subtract this distance from our above result (10.79) for Δx to get the actual length ℓ of the stick in our frame:

$$\ell = \Delta x - \beta^2\Delta x = (1-\beta^2)\Delta x = \left(\frac{1}{\gamma^2}\right)(\gamma\ell') = \frac{1}{\gamma}\ell$$

So in our frame the stick is actually shorter, not longer. This effect is known as length contraction:

> **Length Contraction**: Along its direction of motion, a moving object is shorter than its rest length by a factor of γ.

Note that the effect is only along the direction of motion (our x axis); the two spatial directions perpendicular to this are unaffected. Also note that, like time dilation, length contraction is a property of spacetime itself, not of the moving object: it is space itself that shrinks along the direction of motion.

The Lazy Way to Length Contraction

In a first introduction to length contraction, it is important to work fully through the issues of lack of simultaneity and of the consequent effect of the motion of the rod, as we just did, so that you are aware of and understand these issues. But now that we have been through all that, there is a quicker, easier way to derive the length-contraction effect. Our two events will again be the marking of the two ends of the stick, but this time we will suppose that the stick is at rest in the stationary (unprimed) frame, in which its length is therefore its rest length ℓ and we have $\Delta x = \ell$. We will also suppose that the ends are marked simultaneously in the *moving* (primed) frame, so that $\Delta t' = 0$. We know that the marking of the two ends of the stick will not be simultaneous in the stationary frame, but in that frame the lack of simultaneity doesn't matter: since in the stationary frame the stick is just sitting there at rest, it doesn't matter when its ends are marked in that frame. Using $\Delta x = \ell$ and $\Delta t' = 0$ in the Lorentz transform (10.29), we have

$$\Delta x' = \gamma(\ell - \beta c\Delta t) \tag{10.80a}$$
$$0 = \gamma(-\beta\ell + c\Delta t) \tag{10.80b}$$

Eq. (10.80b) yields $c\Delta t = \beta\ell$, and using this in eq. (10.80a) gives us

$$\Delta x' = \gamma(\ell - \beta(\beta\ell)) = \gamma(1-\beta^2)\ell$$

which, as we saw above, will simplify to

$$\Delta x' = \frac{1}{\gamma}\ell$$

Since it is from the perspective of the moving frame that the stick is in motion, and since the ends of the stick were marked simultaneously in that moving frame, this result for $\Delta x'$ is giving us the length-contraction effect directly.

This alternative derivation of length contraction also illustrates, as would be expected from the principles of relativity, that it does not matter which frame you regard as stationary and which as moving; the physical results, according to the first postulate, will be the same either way. It does not matter whether you work out the length of a moving stick according to a stationary observer or the length of a stationary stick according to a moving observer.

10.4.3 When to Use What

Now that we have the Lorentz transform, time dilation, and length contraction worked out, the question arises, In what sorts of situations do each of these results apply?

The Lorentz transform is a general relation and can be used for any calculation in special relativity.[40] Time dilation and length contraction, however, apply only in certain special circumstances. It is a bit subtle, but if you go back and look at our derivations of those two effects, you can see that you can "get away with" using simple time dilation and length contraction whenever *one of* Δx, Δt, $\Delta x'$, *or* $\Delta t'$ *vanishes.*[41] That is, it's okay to use simple factors of γ rather than the full Lorentz transform as long as, according to one or the other of the two observers, the two events happen at the same time or at the same place. In our derivation of time dilation, for example, the clock was at rest in its own frame, so that we had $\Delta x' = 0$. And in our derivation of length contraction, the two ends of the stick were marked simultaneously in the stick's frame, so that we had $\Delta t' = 0$.

[40] As long, of course, as you use it correctly by taking as your Δt and Δx the time and distance between a pair of events. You cannot, for example, take as Δt the time between the signing of the Magna Carta and the delivery of the Gettysburg Address and as Δx the distance from the rubber to home plate, and then hope to get the corresponding time interval and spatial distance according to alien slime-beings flying by the Earth at three-quarters of the speed of light. Just in case you were inclined to do that. At any rate, this sort of thing has been known to happen in the past, and we hope by mentioning it to save those of you who might otherwise have strayed from the path of righteousness from eternal damnation. Or at least from some embarrassment.

[41] Of course, more than one of these quantities could vanish, but that would leave you with the rather uninteresting case where they *all* vanish.

This also settles the issue of which way around the effect goes — that is, whose time is dilated and whose length is contracted: just as in our derivations of time dilation and length contraction, the observer whose clock runs slower is the one for whom $\Delta x = 0$, and the observer who experiences length contraction is the one for whom $\Delta t = 0$. Although the theory of relativity itself is symmetric — who's moving depends on who you ask, and any situation can be set up equally well from the perspective of either reference frame — , *the situations themselves* can be asymmetric, and it is this asymmetry that dictates which of the frames has the shorter time interval and which the shorter distance.

10.4.4 One Final Note

Lest there be any confusion on this important point, note that the Lorentz transform tells us about the spatial displacements and time intervals between events — the actual displacements and intervals, in the sense of what really *is*. What you would *see* happening is a very different and considerably more complicated question. For the purposes of illustration, let us invoke the classic duo, Alice and Bob: Suppose that a stationary observer Bob just chills out as the moving observer Alice trundles down the x axis toward him. At the moment that Alice passes Bob, they exchange a high five, and we will take this high five as our first event for the purposes of applying the Lorentz transform. After a time $\Delta t'$ passes (according to Alice, who just happens to be looking at her watch), Alice sneezes, which we will take as our second event. If we want to know when Alice sneezes according to Bob, we can use the Lorentz transform relation (10.29b),

$$c\Delta t' = \gamma(-\beta\Delta x + c\Delta t)$$

to figure out Δt: since, between the high five and the sneeze, Alice has been traveling at velocity v for a time Δt according to Bob, she will have traveled a distance $\Delta x = v\Delta t$ further down the x axis. Using this in the above relation, and remembering that $\beta = v/c$ and that $\gamma = 1/\sqrt{1-\beta^2}$, we have

$$\begin{aligned} c\Delta t' &= \gamma(-\beta v\Delta t + c\Delta t) \\ &= \gamma\left(-\beta\frac{v}{c} + 1\right) c\Delta t \\ &= \frac{1}{\sqrt{1-\beta^2}}(-\beta^2 + 1)c\Delta t \\ &= \sqrt{1-\beta^2}\, c\Delta t \end{aligned}$$

and hence

$$\Delta t = \frac{1}{\sqrt{1-\beta^2}}\Delta t' = \gamma\Delta t'$$

This is just the time dilation effect of §10.4.1. It is the time at which Alice sneezes according to Bob, and is correct in every way: if Alice and Bob had synchronized watches as they high-fived, then at the instant Alice sneezes Bob's watch would read whatever $\gamma\Delta t'$ works out to. But how would Bob know that Alice sneezed? He wouldn't. At least not right away. To *see* Alice sneeze, he would have to wait the additional time it would take light to travel from the sneezing Alice back to him. So when you are using the Lorentz transform you should always have in mind that while you are calculating what really and truly actually happens, how you would perceive that that was happening and what you would see are more involved questions.[42]

10.5 The Invariant Interval & Proper Time

In §10.3.4 we saw that the quantity $c^2\Delta t^2 - \Delta x^2$ has the same value in all reference frames, that is, is invariant. This means that regardless of the two frames involved,

$$c^2\Delta t'^2 - \Delta x'^2 = c^2\Delta t^2 - \Delta x^2$$

The square-root of the quantity $c^2\Delta t^2 - \Delta x^2$ is called the invariant interval. The invariant interval is fundamental to relativity — as we saw in §10.3.4 and §10.3.7, it is possible to derive the Lorentz transform and all of special relativity from the postulate that $c^2\Delta t^2 - \Delta x^2$ has the same value in all reference frames — , and for this reason Einstein almost called his theory *invariance theory* rather than *relativity theory.*

When the separation between two events is such that a light pulse traveling from the first event would just make it to the second event, then $\Delta x = c\Delta t$ and consequently $c^2\Delta t^2 - \Delta x^2 = 0$.[43] Events for which $c^2\Delta t^2 - \Delta x^2 = 0$ are said to have a light-like separation and to be on each other's light cones.[44]

If the separation between two events is such that an object traveling at less than the speed of light could make it from the first event to the second, then $\Delta x = v\Delta t < c\Delta t$ and consequently $c^2\Delta t^2 - \Delta x^2 > 0$. Events for which $c^2\Delta t^2 - \Delta x^2$ is positive are said to have a time-like separation. For events

[42] For stills and video clips of relativistic effects as you would see them, you might check out `http://www.anu.edu.au/Physics/Searle/`. There are also some nice Flash apps and animated GIFs at `http://www.phys.unsw.edu.au/einsteinlight/` and `http://www.gmarts.org/index.php?go=420`.

[43] For simplicity, we are assuming here and below that v and hence Δx are in the positive x direction, but since the signs for direction wash out in the squares, the same conclusions about $c^2\Delta t^2 - \Delta x^2$ hold for motion in the negative x direction.

[44] To see why the term *cone* is used we need to work with more than one spatial axis: with x and y axes, $c^2\Delta t^2 - \Delta x^2 = 0$ would become $c^2\Delta t^2 - (\Delta x^2 + \Delta y^2) = 0$ or, equivalently, $\Delta x^2 + \Delta y^2 = c^2\Delta t^2$. If we take the "initial" point in all our Δ's to be the origin ($x = y = t = 0$), this relation simplifies to $x^2 + y^2 = c^2t^2$, which describes a cone around the t axis. With three spatial coordinates, one has a cone in four-dimensional space.

with a time-like separation, it is possible to find a reference frame in which $\Delta x = 0$, that is, in which the two events are separated only by an interval of time Δt, so that they occur at the same place but at different times (whence the term *time-like separation*).

Finally, if the separation between two events is such that even a light pulse could not make it from the first event to the second, then $\Delta x > c\Delta t$ and consequently $c^2\Delta t^2 - \Delta x^2 < 0$. Events for which $c^2\Delta t^2 - \Delta x^2$ is negative are said to have a space-like separation. For events with a space-like separation, it is possible to find a reference frame in which $\Delta t = 0$, that is, in which the two events are separated only by a spatial distance Δx, so that they occur at the same time but at different places (whence the term *space-like separation*).

Only events whose separations are time-like or light-like are causally connected; when the separation of events is space-like, communication between them would require a signal capable of covering the distance that separates them at faster than the speed of light (in fact, at infinite speed, in the frame where the two events occur at the same time but different places). Events whose separation is space-like therefore cannot influence each other and are said to be outside of each other's (event) horizons.

In an object's own reference frame we always have $\Delta x = 0$; in a frame that shares the object's motion, the object is always at rest. In that frame, Δt is the time elapsed on the object's own clock, just as though the object were wearing a wristwatch. This special time is conventionally denoted by τ rather than t and is referred to as the proper time (in the sense that it is the time proper or peculiar to the object itself). Since $\Delta x = 0$, the invariant interval $c^2\Delta t^2 - \Delta x^2$ in this frame reduces to $c\Delta\tau$.

Although in other frames $\Delta x \neq 0$, the squared interval $c^2\Delta t^2 - \Delta x^2$ must, since it is invariant, also work out to $c^2\Delta\tau^2$:

$$c^2\Delta t^2 - \Delta x^2 = c^2\Delta\tau^2 \tag{10.81}$$

If from our perspective the object has a velocity v, then its displacement during time Δt will be $\Delta x = v\Delta t$, so that we have

$$\begin{aligned} c^2\Delta\tau^2 &= c^2\Delta t^2 - \Delta x^2 \\ &= c^2\Delta t^2 - v^2\Delta t^2 \\ &= \left(1 - \frac{v^2}{c^2}\right) c^2\Delta t^2 \\ &= (1 - \beta^2)c^2\Delta t^2 \end{aligned}$$

If we use $1 - \beta^2 = 1/\gamma^2$ from eq. (10.77), this yields

$$c^2\Delta\tau^2 = \frac{1}{\gamma^2}c^2\Delta t^2$$

or

$$\Delta\tau = \frac{1}{\gamma}\Delta t \tag{10.82}$$

which is exactly the time-dilation effect we would expect.

The separations between events — time-like, light-like, and space-like — can be visualized by means of a spacetime diagram. On such diagrams, it is customary to plot x the horizontal axis and to make the vertical axis, not t, but ct, so that it has length dimensions and can be plotted with the same scale as the length dimensions of the x axis. Each point in the plane of the diagram thus corresponds to a point (x, ct) in spacetime. The path of a light ray moving in the positive x direction will then be a straight line of slope

$$\frac{\Delta(ct)}{\Delta x} = c\,\frac{1}{\frac{\Delta x}{\Delta t}} = c\,\frac{1}{c} = 1$$

while that of a light ray moving in the negative x direction will be a straight line of slope -1.

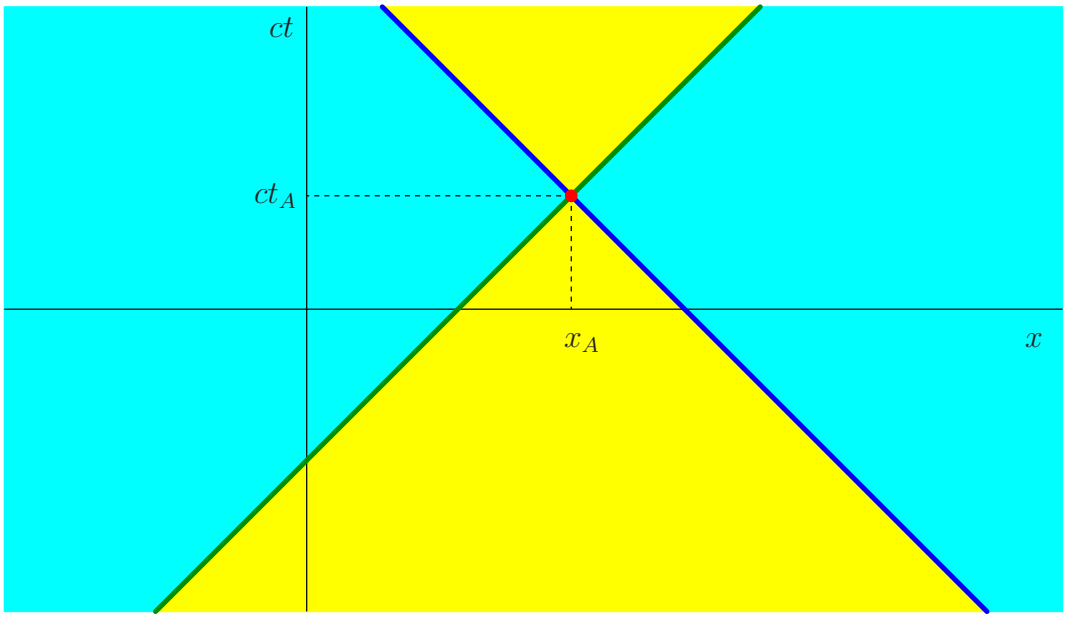


Figure 10.8: What an Awful Color Scheme!

Suppose now we are concerned with some event A that occurs at location x_A at time t_A, corresponding to the point (x_A, ct_A) indicated by the red dot in fig. (10.8). The dark green line, with its slope $+1$, represents the path of a light ray moving in the positive x direction, coming from far out on the negative x axis in the distant past and ending up far out on the positive x axis in the distant future, in such a way that it intersects the point (x_A, ct_A)

corresponding to the event A. In other words, the path of this forward-moving light ray is such that at time t_A it is at x_A. For the spacetime points (x, ct) on the dark green path followed by this light ray,

$$\text{slope} = \frac{x - x_A}{ct - ct_A} = 1$$

which we can rewrite as

$$c^2(t - t_A)^2 - (x - x_A)^2 = 0$$

That is, the separation between event A and any event on the path of this light ray is, as we would have expected, light-like.

Similarly, the dark blue line, with its slope -1, represents the path of a light ray moving in the negative x direction, coming from far out on the positive x axis in the distant past and ending up far out on the negative x axis in the distant future, in such a way that it intersects the point (x_A, ct_A) corresponding to the event A. For the spacetime points (x, ct) on the path of this light ray,

$$\text{slope} = \frac{x - x_A}{ct - ct_A} = -1$$

which can also be rewritten as

$$c^2(t - t_A)^2 - (x - x_A)^2 = 0$$

Once again the separation between event A and any event on the path of the light ray is light-like.

The dark green and dark blue light paths that pass through event A are said to form the light cone of event A. Of course, they don't look like they form much of a cone in our two-dimensional plot with just a single spatial axis x versus time t, but if we also had a y axis oriented perpendicular to the page, the set of light rays passing through event A would indeed form a cone. And with x, y, and z spatial axes, we would have a hypercone in four dimensions. Such light cones are divided into two halves: all those events occurring at times earlier than the vertex event — in this case (x_A, ct_A) — are said to lie on the backward light cone; all those occurring at times later than the vertex event are said to lie on the forward light cone. Events on the backward light cone are such that light coming from them would have just enough time to reach event A; events on the forward light cone are such that light coming from event A would have just enough time to reach them.

Lines connecting event A with points (x, ct) in the yellow regions all have slopes between -1 and $+1$, so that

$$|\text{slope}| = \left|\frac{x - x_A}{ct - ct_A}\right| < 1$$

which we can rewrite as

$$(x - x_A)^2 < c^2(t - t_A)^2$$

or

$$c^2(t - t_A)^2 - (x - x_A)^2 > 0$$

That is, all events that lie inside the light cone of event A have a time-like separation from event A: there is more than enough time for light from events inside the backward light cone to reach A, and more than enough time for light from A to reach events inside the forward light cone.

Similarly, for events outside the light cone of event A,

$$c^2(t - t_A)^2 - (x - x_A)^2 < 0$$

and the separation is space-like: there is not enough time even for light to travel between these events and event A.

And so, to pull all this together, a spacetime diagram plotting x versus ct makes it easy to visualize the relations between events: You draw the light cone for the event with which you are concerned — in our case, event A — by drawing the lines that pass through it with slopes of unit magnitude. All events — that is, all spacetime points (x, ct) — that lie in or on the backward light cone have a time-like or light-like separation from event A and could therefore have causally influenced event A. All events that lie in or on the forward light cone have a time-like or light-like separation from event A and could therefore be causally influenced by event A. And all events that lie outside the light cone are causally disconnected from event A: they cannot have any causal influence on or knowledge of event A and vice versa.

While we won't be doing much with spacetime diagrams, you should have a firm handle on proper time and on the various types of separations between events.

10.6 Addition of Velocities

Suppose an object moving along the x axis has velocity u according to an observer in the unprimed frame.[45] Since we would determine this velocity by the usual displacement/time relation $u = \Delta x/\Delta t$, we can use the Lorentz transform to obtain a result for the velocity u' that the object appears to have in the primed frame:

$$u' = \frac{\Delta x'}{\Delta t'} = \frac{\Delta x'}{\frac{1}{c} \cdot c\Delta t'} = \frac{\gamma(\Delta x - \beta c \Delta t)}{\frac{1}{c}\Big(\gamma(-\beta\Delta x + c\Delta t)\Big)}$$

[45] We use u rather than v for the object's velocity to avoid confusion with the velocity v of the moving frame F' relative to the stationary frame F.

$u =$ velocity of object in frame F
$u' =$ velocity of object in frame F'
$v =$ velocity of F' relative to F

Table 10.1: The Velocities in $u' = \dfrac{u - v}{1 - uv/c^2}$

If we cancel the γ's and multiply the denominator through by the $\frac{1}{c}$, this becomes

$$u' = \frac{\Delta x - \beta c \Delta t}{-\frac{1}{c}\beta \Delta x + \Delta t} \tag{10.83}$$

The object of the game is to get a result for u' in terms of $u = \Delta x/\Delta t$, and we can gerrymander the Δx's and Δt's of eq. (10.83) into the combination $\Delta x/\Delta t$ by dividing through on top and bottom by Δt:

$$u' = \frac{\frac{1}{\Delta t}(\Delta x - \beta c \Delta t)}{\frac{1}{\Delta t}\left(-\frac{1}{c}\beta \Delta x + \Delta t\right)} = \frac{\frac{\Delta x}{\Delta t} - \beta c}{-\frac{1}{c}\beta \frac{\Delta x}{\Delta t} + 1} = \frac{u - \beta c}{-\frac{1}{c}\beta u + 1}$$

We can tidy this up quite a bit by reversing the order of terms in the denominator and replacing β by v/c, which yields

$$u' = \frac{u - \frac{v}{c}c}{1 - \frac{1}{c}\frac{v}{c}u}$$

or

$$u' = \frac{u - v}{1 - uv/c^2} \tag{10.84}$$

This relation is known as the addition of velocities formula; if we know an object's velocity in one reference frame, it allows us to determine that object's velocity in any other reference frame.[46]

When you work with this relation, be careful about which velocity is which: u is the velocity of the object in the unprimed frame F, u' is the velocity of the object in the primed frame F', and v is the velocity of the primed frame relative to the unprimed frame (F' relative to F), as summarized in table (10.1).

The familiar Newtonian result is obtained by taking the limit $u, v \ll c$. In this limit, we have

$$u' \approx \frac{u - v}{1 - 0} = u - v$$

[46] Of course, we are restricting ourselves to the special case that all motion — both of the object and of the reference frames — is along the x axis. A similar result can be obtained for the three-dimensional case, when the object's velocity may be along a different axis from that of the relative velocity of the two frames.

That is, we simply subtract the relative velocity of the observers to determine the velocity of the object in the other frame — just as we did at the beginning of this chapter with the car, the bird, and the camel. But at velocities comparable to the speed of light, the relativistic uv/c^2 term in the denominator has a significant and very important effect: it limits the velocity of the object always to be less than or equal to the speed of light. Consider, for example, the extreme case of a head-on collision between cars both moving at 99% of the speed of light. From our perspective (in the unprimed frame F), car A is moving at $+0.99c$ and car B at $-0.99c$. If we make car A the object and car B the primed frame F', then u is the velocity of car A relative to us and v is the velocity of car B relative to us: $u = 0.99c$ and $v = -0.99c$. From the perspective of someone in car B, the velocity u' with which car A is approaching is therefore

$$u' = \frac{u - v}{1 - uv/c^2} = \frac{0.99c - (-0.99c)}{1 - \dfrac{(0.99c)(-0.99c)}{c^2}} = 0.9999c < c$$

Where we might have expected the combined velocities to seem nearly twice the speed of light, they in fact come out to slightly less then the speed of light.

10.7 Momentum, Energy, & Stuff

In this section, we will develop relativistic expressions for momentum and energy by very physical reasoning. In the following section, we will arrive at these same results in a more elegant but less physically transparent way.

Consider once again the Lorentz transform:

$$c\Delta t' = \gamma(-\beta\Delta x + c\Delta t)$$
$$\Delta x' = \gamma(\Delta x - \beta c\Delta t)$$

The time intervals Δt and $\Delta t'$ and the spatial displacements Δx and $\Delta x'$ are fundamental physical quantities that transform into each other as we go from one reference frame to another: Δt and Δx are combined algebraically and both contribute to each of $\Delta t'$ and $\Delta x'$, so that what is a spatial distance to one observer is partly a time interval to another and what is a time interval to one observer is partly a spatial distance to another. In relativity, we cannot deal separately with spatial distances and time intervals; space and time are inseparably melded together, so that we must instead deal with spacetime displacements that have both spatial and temporal components. While we have been restricting the spatial component to just an x coordinate, were we to extend the Lorentz transform to include the y and z directions by adding

the trivial relations $\Delta y' = \Delta y$ and $\Delta z' = \Delta z$, the sets $(c\Delta t, \Delta x, \Delta y, \Delta z)$ and $(c\Delta t', \Delta x', \Delta y', \Delta z')$ would constitute the components of what are known as four-vectors in a four-dimensional spacetime consisting of three spatial and one temporal dimensions.

Now, the components of a vector are its projections along the axes, and the fact that these components transform in a simple, sensible way is what really, at the most fundamental level, constitutes a vector. For example, a spatial vector in the xy plane has x and y components that transform under rotations in such a way that the components of the rotated vector are the rotated components. While such a vector may of course have vanishing x or y components, a component that equals zero is still component, and a vector in the xy plane must necessarily have both an x and a y component: a vector that lacked, say, a y component wouldn't make any sense, because a rotation of such a vector would automatically generate a y component, leading to a conundrum and logically contradictory nonsense.

In the context of relativity, this means that all four-vectors must have all four spacetime components — not just three spatial components, but also a temporal component. These four components will transform between reference frames according to the Lorentz transform, analogous to the way that spatial vectors in the xy plane transform under rotations. Up to now, we have concerned ourselves only with four-vectors representing spacetime displacements, the components of which are, as noted above, $(c\Delta t, \Delta x, \Delta y, \Delta z)$. We now want to determine the relativistic four-vector corresponding to momentum, and to accomplish this we will follow a line of reasoning very close to that used by Einstein himself when he first formulated relativity.

In the preceding section, we saw that the velocities of objects do *not* transform according to the Lorentz transform between reference frames: while the Δx in the numerator does transform according to the Lorentz transform, so does the Δt in the denominator, so that the velocity $u = \Delta x/\Delta t$ does not; the denominator messes everything up. We therefore look for some nicer definition of velocity that will transform according to the Lorentz transform and that will also reduce to the familiar Newtonian velocity in the nonrelativistic limit. As a first guess, we might try replacing the Δt in $u = \Delta x/\Delta t$ by the proper time interval $\Delta\tau$: $\Delta\tau$, the time that elapses on the object's own clock, is fixed and does not change as we go from one reference frame to another. Now that we have a invariant denominator, our velocity will transform according to the Lorentz transform. Moreover, in the nonrelativistic limit, all observers report the same time interval, so in this limit there is no difference between $\Delta\tau$ and Δt, and consequently between our velocity and the familiar Newtonian velocity. Bingo.

We will thus replace the Newtonian velocity $u = dx/dt$ with [47]

$$u = \frac{dx}{d\tau}$$

and consequently will define the relativistic equivalent of the Newtonian momentum $p = mv$ to be

$$p = mu = m\frac{dx}{d\tau} \tag{10.85}$$

To see how this differs from the Newtonian $p = mv$, we need to re-express the proper time interval $d\tau$ in terms of the time interval dt in our frame. As we found at the end of §10.5, from our perspective the object is experiencing time dilation by a factor of γ:

$$d\tau = \frac{1}{\gamma}dt$$

Therefore

$$p = m\frac{dx}{\frac{1}{\gamma}dt} = \gamma m\frac{dx}{dt} = \gamma mv \tag{10.86}$$

The relativistic momentum thus differs from the Newtonian momentum by a factor of γ.[48] This means that in contrast to the Newtonian momentum's linear growth with velocity, the relativistic momentum, like γ, blows up as we approach the speed of light: $p \to \infty$ as $v \to c$.

$u = dx/d\tau$ was, however, just the x component of the four-vector velocity; the components of the full four-vector are

$$\left(\frac{c\,dt}{d\tau},\ \frac{dx}{d\tau},\ \frac{dy}{d\tau},\ \frac{dz}{d\tau}\right) \tag{10.87}$$

and so the corresponding components of our relativistic four-vector momentum are

$$\left(m\frac{c\,dt}{d\tau},\ m\frac{dx}{d\tau},\ m\frac{dy}{d\tau},\ m\frac{dz}{d\tau}\right) \tag{10.88}$$

[47] We are now switching from the finite-interval Δ notation that we have heretofore predominantly been using to differentials dt and $d\tau$, but of course these differentials transform under the Lorentz transform the same way that finite intervals Δt and $\Delta\tau$ do, so this need not cause any confusion.

[48] In the early days of relativity, as a result of some experimental relations found by W. Kaufmann between the velocity and apparent mass of an electron, some people liked to associate the γ closely with the m and called γm the "relativistic mass," as though the object became more massive the faster it was moving. In fact, mass is a fundamental physical quantity that is invariant under the Lorentz transform; it does not change between reference frames or with the speed of the object, and to think it does is, well, just plain wrong. But if you see the term "relativistic mass" in a book, that's what it's referring to.

The y and z components of this momentum are just like the x component; the more interesting question is what its time component corresponds to physically. To answer this question, we simply follow our noses: again using $d\tau = dt/\gamma$, we have

$$m\frac{c\,dt}{d\tau} = m\frac{c\,dt}{\frac{1}{\gamma}dt} = \gamma mc \tag{10.89}$$

It's not immediately obvious how to interpret γmc. But in rummaging around for some way to make progress, we remember that relativistic results for quantities should, in the limit of velocities far less than the speed of light, reduce to the corresponding Newtonian results. This suggests that to see to what Newtonian quantity γmc corresponds we might try looking at its low-velocity limit. To do this, we first need to write γmc in terms of the velocity v:

$$\gamma mc = \frac{1}{\sqrt{1 - v^2/c^2}}\, mc$$

To see more clearly how this behaves when $v \ll c$, we can do a Taylor expansion of γ:

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} = 1 + \frac{1}{2}\frac{v^2}{c^2} + \frac{3}{8}\frac{v^4}{c^4} + \frac{5}{16}\frac{v^6}{c^6} + \cdots \tag{10.90}$$

where the omitted terms involve successively higher powers of v/c. For velocities much less than the speed of light, v^2/c^2 is already very small, and by comparison we can neglect the v^4/c^4 and higher terms. In the nonrelativistic limit, γmc thus reduces to

$$\gamma mc \approx \left(1 + \frac{1}{2}\frac{v^2}{c^2}\right) mc = mc + \frac{\frac{1}{2}mv^2}{c}$$

Comparing the second term with the Newtonian kinetic energy $\frac{1}{2}mv^2$ leads us to conclude that the time component of the momentum is therefore energy over the speed of light, so that $E/c = \gamma mc$, or

$$E = \gamma mc^2 \tag{10.91}$$

This is the relativistic result for energy and the origin of the popularly quoted special case $E = mc^2$. There are several important things to note about this result:

- Since
$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} \to \infty \quad \text{as} \quad v \to c$$
it would take an infinite quantity of energy to accelerate a mass all the way up to the speed of light. Massive objects are therefore constrained

always to move at speeds slower than the speed of light. For small masses like elementary particles, you can get really close to the speed of light — particles at the big colliders are routinely accelerated to 99.9999...% of the speed of light — , but you can never get all the way up to the speed of light. This is entirely different from the Newtonian $\frac{1}{2}mv^2$, which has no upper limit on v.

- In the nonrelativistic limit (that is, for velocities far less than the speed of light), we have, as we worked out above,

$$E \approx mc^2 + \tfrac{1}{2}mv^2$$

For low velocities, the velocity-dependent part of the energy is therefore, as it should be, the same as the familiar Newtonian $\frac{1}{2}mv^2$. But there is also the constant term mc^2, which is present even when $v = 0$, and which is therefore called the rest energy of the mass. This term leads us to suspect that it might be possible to convert mass into energy and vice versa, and in fact conversions in both directions are routine: nuclear reactors and nuclear weapons convert mass into energy, and at the big particle colliders, the energy of the colliding particles is used to create mass, in the form of exotic new particles. Because the speed of light c is so large in everyday units (3.0×10^8 m/s), the conversion of even a small quantity of mass yields a huge quantity of energy.

- $E = \gamma mc^2$ includes the rest energy mc^2 of the mass. The relativistic equivalent of kinetic energy — that part of E that depends only on the motion of the mass — may be obtained by subtracting out this rest energy. Thus the relativistic kinetic energy is

$$K = E - mc^2 = \gamma mc^2 - mc^2 = (\gamma - 1)mc^2 \tag{10.92}$$

Recall now our discussion of travel to distant stars in §10.4.1: in order for it to be humanly possible to reach such stars, a great deal of time dilation and hence a large γ would be needed. We would, for example, need $\gamma = 1000$ to reduce the travel time to a star 1000 light-years away to about one year. But according to $K = (\gamma - 1)mc^2$, the energy required to boost a rocket ship to such a speed would be $999mc^2$, that is, 999 times the rest energy of the rocket. The most effective way to produce large quantities of energy is by the conversion of mass, but even with 100% efficiency this would still require us to convert a mass 999 times larger than that of the rocket itself completely into energy. The production of such a large amount of energy is vastly beyond our current capabilities and seems very unlikely ever to be within our reach. We can't even figure out how to power our suburban assault vehicles and microwave ovens without destroying the Earth; fat chance we're ever going

to manage space travel. But, disappointing as this may seem, it's probably not a bad thing, considering the way we behave ourselves and what we've done to our own planet.

Even if we can't travel to distant star systems, we can, with a little gerrymandering, get another useful relation involving momentum and energy. Recall that the interval $c^2\Delta t^2 - \Delta x^2$ was invariant under the Lorentz transform, so that it had the same value in all reference frames. Consider the corresponding combination with momentum and energy:

$$p^2_{temporal} - p^2_{spatial} = \left(\frac{E}{c}\right)^2 - p^2 = \left(\frac{\gamma mc^2}{c}\right)^2 - (\gamma mv)^2 = \gamma^2\left(1 - \frac{v^2}{c^2}\right) m^2c^2$$

Recalling now, from eq. (10.77), that

$$1 - \frac{v^2}{c^2} = 1 - \beta^2 = \frac{1}{\gamma^2}$$

we have

$$\left(\frac{E}{c}\right)^2 - p^2 = m^2c^2$$

which, if we tidy up by multiplying through by c^2, becomes

$$E^2 - p^2c^2 = m^2c^4 \tag{10.93}$$

Although we won't ourselves have much practical use for it, eq. (10.93) is a very important and fundamental relation between energy and momentum. The advantage of eq. (10.93) over the relations $E = \gamma mc^2$ and $p = \gamma mv$ is that it applies to massless as well as massive objects. For a massless particle, eq. (10.93) tells us that $E = pc$ — which, if you think about $E = \gamma mc^2$ and $p = \gamma mv$, means that $v = c$:

$$\lim_{m\to 0} \frac{E}{p} = \lim_{m\to 0} \frac{\gamma mc^2}{\gamma mv} = \frac{c^2}{v}$$

so that $E/p = c$ means that $c^2/v = c$ and hence that $v = c$. Massless objects are thus constrained to move at exactly the speed of light; they can never slow down (or speed up). Conversely, any object that moves at the speed of light must be massless.

Since the relativistic momentum four-vector, like any other four-vector, transforms according to the Lorentz transform, the spatial component p'_x and temporal component $p'_t = E/c$ in a frame F' will be related to the corresponding components p_x and p_t in a frame F by

$$\begin{aligned} p'_x &= \gamma(p_x - \beta p_t) \\ p'_t &= \gamma(-\beta p_x + p_t) \end{aligned} \tag{10.94}$$

Because momentum (in the sense of the spatial component p_x of the four-vector momentum) and energy ($E = p_t c$) are thus transformed into each other, so that the energy in one frame contributes to the momentum in another and vice versa, in relativity it is not possible to conserve just momentum and not also energy (as we did, for example, in inelastic Newtonian collisions): if momentum but not energy were conserved in one frame, we could always Lorentz transform into another frame in which the nonconservation of $p_t = E/c$ would, by the former of eqq. (10.94), result in the nonconservation of p'_x. So in relativity both the momentum and the energy of an isolated system are always conserved. Note, however, that this does not mean that all collisions are elastic in relativity: just as in Newtonian physics, conversions between kinetic and the various other forms of energy can occur in relativistic collisions. In particular, mass can be converted into energy (as is done in nuclear weapons and power plants) and vice versa (as is done when the energy of colliding particles produces heavier, more exotic particles in particle colliders). And in Newtonian collisions the energy of an isolated system is really also always conserved, if you include all of its various forms; the mechanical kinetic energy that is lost in collisions is still there in the microscopic form of heat or in the form of potential energies.

Finally, for those of you who were curious about the relativistic expression for force, there are various ways to show that the relativistic four-vector force K is, as you would expect, given by[49]

$$K = \frac{dp}{d\tau}$$

This relation is analogous, and in the nonrelativistic limit reduces, to the Newtonian

$$F = \frac{dp}{dt}$$

$K = dp/d\tau$ is thus the equivalent of Newton's second law, $F = ma$. As applying this relation is, however, much more involved than applying $F = ma$ and would be neither terribly illuminating nor of practical use to us, we will not pursue it any further.

10.7.1 A Nicer Derivation of Momentum & Energy

We can also arrive at the relativistic expressions for momentum and energy by considering the simple special case of a mass at rest, giving this mass a velocity by applying the Lorentz transform, and then requiring that the result we so obtain reproduce the familiar Newtonian momentum in the nonrelativistic limit. As a four-vector, the relativistic momentum p should, as

[49] K, rather than F, is the conventional notation for relativistic forces.

argued in the previous section, have both a temporal component p_t and a spatial component p_x. Although at first we have no idea what either of these components should be, we do expect that the spatial component p_x should vanish when the object is at rest, so that in the object's own reference frame F we should have $(p_t, p_x) = (p_t, 0)$.

Recall now that the inverse Lorentz transform (10.36) of p.537,

$$\begin{aligned} \Delta x' &= \gamma(\Delta x + \beta c \Delta t) \\ c\Delta t' &= \gamma(\beta \Delta x + c\Delta t) \end{aligned}$$

will take us from the object's own reference frame F to a frame F' moving in the negative x direction at speed $-v$, which is just what we want: in this new frame the object will appear to be moving in the positive x direction at speed $+v$. So if we apply this transform to our momentum $(p_t, p_x) = (p_t, 0)$ for a stationary mass, we will have the momentum (p'_t, p'_x) of a mass moving at velocity $+v$:

$$p'_x = \gamma(p_x + \beta c p_t) = \gamma(0 + \beta c p_t) = \gamma \beta c p_t \tag{10.95a}$$

$$c p'_t = \gamma(\beta p_x + c p_t) = \gamma(0 + c p_t) = \gamma c p_t \tag{10.95b}$$

We want this spatial component of the relativistic momentum p'_x to match up with the Newtonian mv for $v \ll c$. If we simplify it a bit by

$$p'_x = \gamma \beta c p_t = \gamma \frac{v}{c} c p_t = \gamma v p_t$$

and note that in the nonrelativistic limit

$$\gamma = \frac{1}{\sqrt{1 - v^2/c^2}} \approx 1$$

then we see that for $v \ll c$ we have

$$p'_x \approx v p_t$$

This will reproduce the Newtonian mv if $p_t = m$. We thus conclude that the spatial component of the relativistic momentum is $\gamma v p_t = \gamma m v$, just as in eq. (10.86). And the temporal component of the relativistic momentum is, from eq. (10.95b),

$$c p'_t = \gamma c p_t = \gamma m c$$

just as we obtained in eq. (10.89). From here, the interpretation of the quantity γmc as E/c proceeds exactly as in the preceding section.

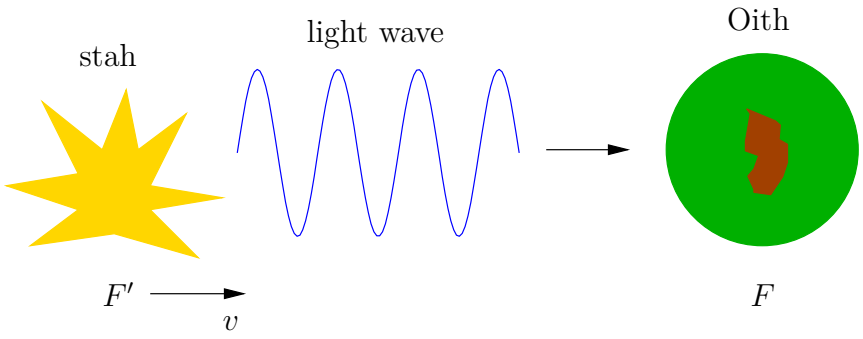


Figure 10.9: Light Coming to New Jersey from a Star in the Cannabis System

10.8 The Doppler Shift

Consider a light wave traveling to us from some star. We will regard ourselves on Earth as the stationary frame F and the star, which we will take to be moving toward us at velocity v, as the moving frame F', as shown in fig. (10.9). We want to relate the frequency of the starlight according to its source (the star) to its frequency according to the observer (us here on Earth). For this purpose, it will be easiest to think of the light wave as though it were simply passing by the star on its way to us and to take as the two events to which we will apply the Lorentz transform the passage of two successive wave crests. In the star's frame F' these two crests are observed at the same place, so that $\Delta x' = 0$, and the time $\Delta t'$ between crests is the period T' of the wave according to the star. Applying the inverse Lorentz transform (10.36) from p.537 with $\Delta x' = 0$ and $\Delta t' = T'$, we have [50]

$$\Delta x = \gamma(\Delta x' + \beta c \Delta t') = \gamma \beta c T'$$
$$c\Delta t = \gamma(\beta \Delta x' + c\Delta t') = \gamma c T'$$

Because $\Delta x \neq 0$, $\Delta t = \gamma T'$ is *not* the period of the wave in our frame; from our perspective, the star was moving forward during the time between wave crests, so that the trailing wave had to cover the extra distance $\Delta x = \gamma \beta c T'$ in order to catch up with the star. Since the wave is traveling at the speed of light according to us (as of course it is according to any observer), the time it took the wave to catch up according to us is

$$t_{catch\ up} = \frac{\Delta x}{c} = \frac{\gamma \beta c T'}{c} = \gamma \beta T'$$

[50] Note that we could have obtained the latter of these results immediately by observing that since $\Delta x' = 0$, simple time dilation applies and therefore $\Delta t' = \Delta t / \gamma$. The former of these results would then follow from $\Delta x = v\Delta t = v\gamma \Delta t' = \frac{v}{c} c\gamma \Delta t' = \beta c \gamma \Delta t'$.

To extract our result for the period of the wave from Δt, we need to subtract out this extra time the trailing wave crest spent catching up with the star. Thus

$$T = \Delta t - t_{catch\ up} = \gamma T' - \gamma\beta T' = \gamma(1-\beta)T'$$

If now we recall that by definition $\gamma = 1/\sqrt{1-\beta^2}$, this can be simplified to

$$T = \sqrt{\frac{1}{1-\beta^2}}\,(1-\beta)T' = \sqrt{\frac{1}{(1-\beta)(1+\beta)}}\,(1-\beta)T' = \sqrt{\frac{1-\beta}{1+\beta}}\,T'$$

If we use $T = 1/f$ to relate the period T to the frequency f of the wave, this becomes

$$\frac{1}{f} = \sqrt{\frac{1-\beta}{1+\beta}}\,\frac{1}{f'}$$

which, when inverted to give f (the frequency we observe) in terms of f' (the frequency the star emitted), yields

$$f_{observed} = f_{emitted}\sqrt{\frac{1+\beta}{1-\beta}} \tag{10.96}$$

Eq. (10.96) is the Doppler Shift for light.[51] Be mindful that eq. (10.96) is specifically for the case that the source (the star) is approaching the observer (us) at velocity $v = \beta c$; if the source is receding (moving away) from the observer, then v and β will have negative values.

For example, if a star is receding from us at half the speed of light, then $\beta = -\frac{1}{2}$ and

$$f_{observed} = f_{emitted}\sqrt{\frac{1+(-\frac{1}{2})}{1-(-\frac{1}{2})}} = f_{emitted}\sqrt{\frac{1}{3}} = 0.577\ f_{emitted}$$

The starlight is thus shifted to a lower frequency, which, since $c = \lambda f$, means it is shifted to a longer wavelength. Although neither the emitted nor observed frequency and wavelength need be in the visible part of the electromagnetic spectrum, since the longest visible wavelengths are red, light that is Doppler-shifted to a longer wavelength is said to be red-shifted. Likewise, since the shortest visible wavelengths are blue, light Doppler-shifted to a shorter wavelength is said to be blue-shifted.

Newton, like a lot of insecure people, wanted to believe that the universe was infinite, static, and eternal. It turns out, however, that the gravitational field equations of general relativity predict that the universe is either expanding or contracting.[52] Oddly enough, such solutions to the field equations were first put forth by the Russian mathematician A.A. Friedmann:

[51] As opposed to the Doppler Shift for sound, covered in §9.8.4.

[52] A fuller, more mathematical treatment of the expansion of the universe is presented in §22.8.1.

Einstein had realized before Friedmann's work that the field equations predicted an expanding universe, but, like Newton, he wanted to believe that the universe was eternal and static. Mathematically, it was possible to modify the field equations by the addition of a term proportional to the spacetime metric; the constant of proportionality is called the cosmological constant. By allowing for a nonzero cosmological constant, Einstein was able to obtain solutions corresponding to a static universe. Shortly afterward, however, the astronomer Edwin Hubble ascertained experimentally that the universe is in fact expanding exactly as predicted by the Friedmann's solutions: the farther a star is from us, the more red-shifted its light is, and hence the faster it is receding from us. It is as though the stars were dots on the skin of a balloon that is being inflated: as the balloon gets bigger, the dots get farther apart, and the rate at which they are getting farther apart is proportional to how far they are from each other. Einstein was forced to concede that he had erred and later considered his introduction of a nonzero cosmological constant the biggest mistake of his life.[53] It turns out, however, that he spoke too soon: we now know that there is a kind of matter known as a Higgs field that, in addition to giving matter its mass, makes a nonzero contribution to the cosmological constant. This contribution produced an exponential expansion of the early universe known as inflation, a mechanism central to modern cosmology.[54]

10.9 General Relativity

What we have been studying so far is known as *special relativity* because the postulate that physics is the same in all inertial frames is a special case: the case of inertial (nonaccelerating) reference frames. After formulating special relativity, Einstein wanted to make the theory even simpler (and therefore more powerful and beautiful) by doing away with this restriction and postulating that physics is the same in all frames, whether inertial or not. Because this new theory is no longer restricted to inertial frames, it is known as *general relativity*.

As a simple example of the kinds of effects one would expect when dealing with accelerating reference frames, consider a rotating disk: by virtue of their circular motion, the various points on the disk are experiencing a centripetal acceleration. If the disk were not rotating, the relation between the circumference C and radius r would be the prosaic $C = 2\pi r$, as shown on the left

[53] Actually, there is apparently some debate about the degree of regret he felt.

[54] The precise identity of the inflaton particle responsible for inflation has actually not yet been determined. But while it would be nice to discover it, there is no shortage of candidates in modern theories, and the mechanism of inflation is soundly established even without the inflaton having been precisely identified. For more about inflation, see §22.8.

Figure 10.10: Still Life with Disk and Bowl

side of fig. (10.10), with the circumference in red and the radius in blue just to make things more exciting. When the disk is rotating, however, we expect that the points on the perimeter of the disk will, by virtue of their tangential velocity, be experiencing length contraction and that therefore $C < 2\pi r$. One way to make sense of this is to think of the acceleration of rotating disk as corresponding to a curvature of space, such that what was a planar disk in some sense takes on the geometry of a bowl, as shown on the right side of fig. (10.10): with the blue radius measured along the surface of the bowl, we will indeed have the red circumference C around the rim less than $2\pi r$.

This sort of reasoning suggests that in order to generalize relativity to take into account noninertial, accelerating reference frames, one should think in geometric terms, and this turns out to require some rather sophisticated math called differential geometry. It took Einstein himself eleven years to fully formulate the new theory. But as an added bonus it turns out that general relativity is much more than just an extension of special relativity to the case of noninertial frames: built into general relativity, by its very nature, is a theory of gravity.

In Newtonian mechanics, the inertial mass in $F = ma$ is assumed to be the same as the gravitational mass on which Newton's law of gravity acts, that is, the same as the m_1 and m_2 in

$$F = \frac{Gm_1m_2}{r^2}$$

Gravitational and inertial mass have been experimentally verified to be equivalent out to many orders of magnitude, but there is no theoretical reason in Newtonian mechanics why they have to be equal to each other. The extension of relativity to noninertial reference frames, however, requires that any acceleration you experience be physically identical, whether gravitational or inertial, so that the physics you experience way out in empty space in a spaceship accelerating at one g is identical to the physics you experience while standing at rest on the Earth under the influence of its one g of grav-

ity. This principle of equivalence thus *requires* that gravitational and inertial mass in principle be identical.

Unfortunately the math required to do anything quantitative with general relativity is far beyond what we can get into, but some of its more significant qualitative aspects are: [55]

- In a rather abstract mathematical sense, the presence of mass (or, equivalently, energy) curves spacetime and determines its geometry. This curvature of spacetime expresses itself as the gravitational force, and objects in free-fall follow the shortest routes through this curved spacetime.[56] These shortest routes are known as geodesics, and for weak gravitational fields (like those we experience everyday on Earth and those that govern the motion of the planets) they match the trajectories predicted by Newton's law of gravity.

- Gravity causes time dilation, so that, just by being in the Earth's gravitational field, time is passing (very slightly) more slowly for you than for someone way out in inter-stellar space. The stronger the gravitational field, the greater the dilation. This effect has been verified experimentally by measuring the change in the frequency (the red-shift) of light waves rising through the Earth's gravitational field.[57]

- Light is massless (which is why it travels at the speed of light).[58] Light does, however, have energy, and, energy being equivalent to mass, light will therefore feel the pull of gravity and be deflected by gravitational fields just like any other form of matter. The amount of deflection is small, but it has been verified experimentally by measuring the bending of starlight that grazes the Sun: Since the brightness of the Sun would otherwise make the starlight impossible to see, one has to wait for an eclipse. A photograph is then taken of the stars adjacent to the Sun and compared to a photograph of the same region of the sky when the Sun is not between it and the Earth. When the photographs are overlaid, the bending of a star's light as it passes very close to the Sun is visible as a shift in the apparent location of that star relative to the other stars, as shown in fig. (10.11).

[55] In the following sections we do, however, give a mathematical outline of the field equations of general relativity and of some of the effects listed here, just so that those of you who are curious can get the gist of them.

[56] On a sphere, for example, the shortest route from the north to the south pole is a semicircular arc along a line of longitude, and more generally the shortest route between two points on the surface is along the great circle that passes through the points.

[57] For a quantitative treatment of this effect, see §10.9.2.

[58] Amazing how these things work out, isn't it? ☺

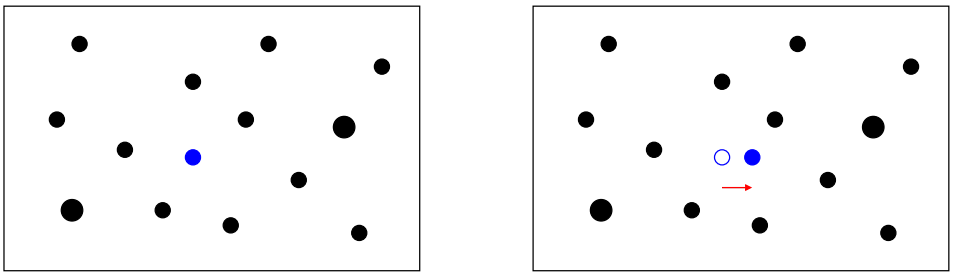


Figure 10.11: A Shifty Star

- In general relativity, the orbits of the planets are not quite elliptical. In fact, they are not even closed; the planets, upon completing a full revolution, do not come back to quite the same place from which they started out. As a result, their perihelia precess — which is a fancy way of saying that their very nearly elliptical orbits seem themselves to rotate, as more or less shown in fig. (10.12). This effect, though a minuscule 43 seconds of arc of rotation per century even for a rapidly revolving planet the like Mercury, was actually observed and measured before it was predicted by general relativity, and the measurements are in agreement with the theory.

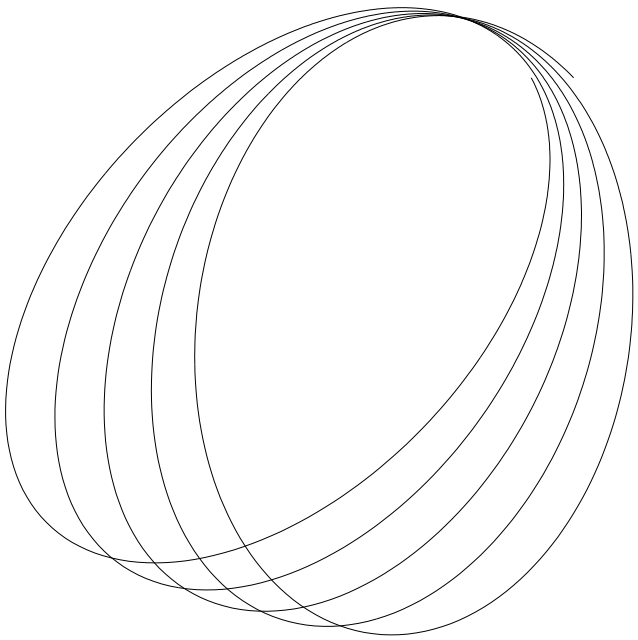


Figure 10.12: Precession of a Perihelion

- And, what seems — for reasons we've never been able to fathom — an endless source of fascination for the general public, general relativity allows for the existence of black holes,[59] aggregations of mass so dense that not even light can escape from them. Black holes must therefore be detected indirectly, by the radiation given off by matter falling into them, the bending of light that passes by them, or their gravitational effect on other, visible masses. At this point there is, however, sound agreement among astronomers that black holes not only have been observed but are fairly common, being found at the centers of most galaxies (including our own Milky Way).

10.9.1 The Field Equations

The big-people notation for the four-vector (ct, x, y, z) is x^μ, where the index μ runs from 0 to 3, with x^0 corresponding to ct and x^1, x^2, x^3 to x, y, z. Derivatives with respect to these coordinates are denoted by ∂_μ:

$$\partial_\mu = \frac{\partial}{\partial x^\mu}$$

According to what is known as the summation convention, indices that are repeated within a term are understood to be summed over. Thus, for example, $\partial_\mu x^\mu$ stands for

$$\begin{aligned}\partial_\mu x^\mu &= \sum_{\mu=0}^{3} \partial_\mu x^\mu \\ &= \frac{\partial(ct)}{\partial(ct)} + \frac{\partial x}{\partial x} + \frac{\partial y}{\partial y} + \frac{\partial z}{\partial z} \\ &= 1 + 1 + 1 + 1 \\ &= 4\end{aligned}$$

Also, when we write $f(x)$, it means that f is a function of the four-vector spacetime location x, that is, of (ct, x, y, z).

In general relativity, all of the physics is determined by the geometry of spacetime, and this geometry is determined mathematically by the metric (tensor) $g_{\mu\nu}(x)$. The metric $g_{\mu\nu}$ is defined by the relation

$$ds^2 = dx^\mu \, g_{\mu\nu} \, dx^\nu \tag{10.97}$$

[59] Although a black-hole solution to the field equations of general relativity was worked out by Karl Schwarzschild in 1916, the term *black hole* was not coined until 1967, by John Wheeler. While the term quickly came to be used universally, initially it met with strong objections from some French and Russian physicists; it seems that its literal translation into their languages is a nasty slang term for a part of the body typically held in very low esteem.

where dx is an infinitesimal spacetime displacement and ds is the corresponding infinitesimal distance traveled. This relation is the generalization of the Pythagorean theorem to curved spaces. In the flat space of an xy plane, for example, we have

$$ds^2 = dx^2 + dy^2 = \begin{pmatrix} dx & dy \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} dx \\ dy \end{pmatrix}$$

so that the metric is simply

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

If this same plane is instead parametrized in polar coordinates, we have

$$ds^2 = dr^2 + r^2\, d\theta^2 = \begin{pmatrix} dr & d\theta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \begin{pmatrix} dr \\ d\theta \end{pmatrix}$$

so that the metric is now[60]

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix}$$

More precisely, in relativity ds is the invariant interval, that is, $ds = c\, d\tau$, where τ is the proper time. Thus in special relativity we have

$$\begin{aligned} ds^2 &= c^2\, d\tau^2 \\ &= c^2\, dt^2 - (dx^2 + dy^2 + dz^2) \\ &= \begin{pmatrix} c\, dt & dx & dy & dz \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} c\, dt \\ dx \\ dy \\ dz \end{pmatrix} \end{aligned}$$

so that the metric

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

This metric of special relativity is known as the flat-space metric because the geometry to which it corresponds, while non-Euclidean, has no curvature.

[60] We should perhaps note, however, that polar coordinates, while providing a nicely simple example for our purposes, actually constitute what is termed a non-coordinate basis and would therefore be unnatural for calculations in general relativity. Natural are coordinate bases, which are related to one another by general coordinate transforms.

All of the geometrical properties of a spacetime can be constructed from the metric $g_{\mu\nu}$, and it turns out that the most general construct you can build from it is the Riemann curvature tensor

$$R^{\lambda}{}_{\mu\nu\kappa} = \partial_\kappa \Gamma^{\lambda}_{\mu\nu} - \partial_\nu \Gamma^{\lambda}_{\mu\kappa} + \Gamma^{\rho}_{\mu\nu}\Gamma^{\lambda}_{\kappa\rho} - \Gamma^{\rho}_{\mu\kappa}\Gamma^{\lambda}_{\nu\rho}$$

where Γ is the affine connection

$$\Gamma^{\lambda}_{\mu\nu} = \tfrac{1}{2} g^{\lambda\kappa}\Big(\partial_\mu g_{\kappa\nu} + \partial_\nu g_{\kappa\mu} - \partial_\kappa g_{\mu\nu}\Big)$$

The Ricci tensor $R_{\mu\nu}$ is the contracted form of the Riemann curvature,

$$R_{\mu\nu} = R^{\lambda}{}_{\mu\nu\lambda}$$

In his belief that the correct theory is the one that is "as simple as possible but not simpler," Einstein asked himself what would be the simplest general relation between the curvature of spacetime and matter that would reproduce the physics we actually see in the universe. The mathematical answer to this is the field equations of general relativity:

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R = -\tfrac{8\pi G}{c^4} T_{\mu\nu} \tag{10.98}$$

where G is the familiar universal gravitational constant and $T_{\mu\nu}$ is the energy-momentum tensor, which corresponds to the distribution of matter and energy that give rise to the gravitational field. It is possible, as Einstein later did, to modify these field equations by including a term with a cosmological constant Λ:[61]

$$R_{\mu\nu} - \tfrac{1}{2} g_{\mu\nu} R + \tfrac{1}{c^2}\Lambda g_{\mu\nu} = -\tfrac{8\pi G}{c^4} T_{\mu\nu}$$

It is also possible to formulate field equations involving terms with higher derivatives. Field relations involving such higher-derivative terms, which in the nonrelativistic limit still reproduce the physics of the universe we see around us, may be a natural consequence of string and other more comprehensive theories.

The object of the game when doing calculations in general relativity is to solve for the metric corresponding to a given distribution of matter and energy $T_{\mu\nu}$. This is very difficult to do because the field equations are highly nonlinear: the curvature of spacetime, which constitutes the gravitational field, itself contributes to the energy density that gives rise to the curvature. One basic solution to the field equations is the Schwarzschild metric, which corresponds to the static gravitational field outside of a spherically symmetric gravitational distribution of matter. If one uses spherical coordinates for the

[61] For more on the cosmological constant and it's effect, see §22.8.1.

spatial coordinates, the Schwarzschild metric can be expressed as

$$g_{\mu\nu} = \begin{pmatrix} 1 - \dfrac{2mG}{rc^2} & 0 & 0 & 0 \\ 0 & -\left(1 - \dfrac{2mG}{rc^2}\right)^{-1} & 0 & 0 \\ 0 & 0 & -r^2 & 0 \\ 0 & 0 & 0 & -r^2 \sin^2\theta \end{pmatrix} \tag{10.99}$$

where m is the mass giving rise to the gravitational field. As you can see, there is a singularity at the black-hole radius

$$r = \frac{2mG}{c^2}$$

Because of the c^2 in the denominator and the tiny value of G, the density of matter required to produce a black hole is huge: to make the Earth into a black hole, you would, as you can see by substituting in the values of G, c, and the Earth's mass, have to compress it down to a radius of a little less than a centimeter.

Another complication of general relativity is that the correspondence between metrics and spacetime geometries is not one-to-one: it is the geometry that is physically meaningful, but because the metric does not depend on the coordinate system used to express it, any given geometry corresponds to an infinite number of metrics. We saw this above in the case of the metric for an xy plane in the case of Cartesian and polar coordinates. Indeed, the whole point of general relativity is that physics is the same in all reference frames, that is, it doesn't matter what coordinate system you work with. Metrics that are geometrically equivalent are said to be connected by a general coordinate transform. Invariance under general coordinate transforms is in fact the symmetry that gives rise to gravity. This is problematic for the formulation of a quantum theory of gravity by path integrals, which involves a functional integral like

$$\int Dg \ \exp\left(-\frac{i}{16\pi G}\frac{c^3}{\hbar}\int d^4x\, R(x)\sqrt{-\det g(x)}\right)$$

where the Dg means that the integration is over all possible geometries, that is, over all possible metrics g that are not connected by a general coordinate transform. At present, no one knows how to carry out such an integration; for technical reasons, methods that work for other quantum theories do not work for gravity.[62]

[62] Path integrals are only one approach to quantizing gravity — there are several — , but all present some form of as yet insurmountable mathematical difficulty.

10.9.2 Gravitational Time Dilation

Although we are not in a position to derive the Schwarzschild metric (10.99), we can, taking it as a given, show how time slows down in a gravitational field. For a stationary object, only the time coordinate changes, so that

$$dx^\mu = (c\,dt, dx, dy, dz) = (c\,dt, 0, 0, 0)$$

In other words, only $dx^0 = c\,dt$ is nonzero. With the Schwarzschild metric, the value of ds^2 in eq. (10.97) is thus

$$ds^2 = dx^\mu\, g_{\mu\nu}\, dx^\nu = dx^0\, g_{00}\, dx^0 = g_{00}\,(dx^0)^2 = \left(1 - \frac{2mG}{rc^2}\right) c^2\, dt^2$$

Now, just as it was in special relativity, this ds^2 is the square of the proper time interval: $ds^2 = c^2\, d\tau^2$. Using $ds^2 = c^2\, d\tau^2$ in the above relation and dividing out a c^2, we therefore arrive at a relation between the proper time interval $d\tau$ — the time elapsed on a stationary clock — and the change dt in the time coordinate t by which we are parametrizing everything:

$$d\tau^2 = \left(1 - \frac{2mG}{rc^2}\right) dt^2 \tag{10.100}$$

In particular, for a clock infinitely far from the mass m, we have, taking $r \to \infty$,

$$d\tau_\infty^2 = dt^2 \tag{10.101}$$

This clock out at infinity will, being outside of m's gravitational field, tick at what we would regard as a "normal" rate — the rate a clock would tick at in the absence of gravitational effects. Taking the ratio of eqq. (10.100) and (10.101), we have

$$\left(\frac{d\tau}{d\tau_\infty}\right)^2 = \left(1 - \frac{2mG}{rc^2}\right)$$

or

$$d\tau = d\tau_\infty \sqrt{1 - \frac{2mG}{rc^2}}$$

From this result, we see that the smaller r, the smaller the time $d\tau$ elapsed relative to a clock outside the gravitational field. That is, the stronger the gravitational field, the more slowly time elapses. At the black-hole radius $r = 2mG/c^2$, $d\tau$ vanishes and time stands still.[63]

[63] This does not, however, mean that someone falling into a black hole never reaches the black hole. It is only from our perspective as observers outside the black hole that the person falling into it would never reach it — they would seem to us to be moving more and more slowly as they approached the surface of the black hole — ; from the perspective of the person falling into the black hole, time would seem to be elapsing at a normal rate, and in fact the integral of the changes $d\tau$ in this person's proper time yields a finite result for the time to reach the black hole.

For a clock at the surface of the Earth, with

$$m \approx 5.97 \times 10^{24}\,\text{kg} \qquad G \approx 6.67 \times 10^{-11}\,\text{Nm}^2/\text{kg}^2$$
$$r \approx 6.38 \times 10^{-11}\,\text{m} \qquad c \approx 3.00 \times 10^{8}\,\text{m/s}$$

we have

$$\begin{aligned} d\tau &= d\tau_\infty \sqrt{1 - \frac{2(5.97 \times 10^{24})(6.67 \times 10^{-11})}{(6.38 \times 10^{6})(3.00 \times 10^{8})^2}} \\ &= d\tau_\infty \sqrt{1 - 1.39 \times 10^{-9}} \\ &\approx d\tau_\infty (1 - 7 \times 10^{-10}) \end{aligned}$$

That is, a clock on the surface of the Earth would, because of gravitational effects, tick slower than a clock far out in empty space by about one part in 7×10^{-10}, or about $\frac{1}{50}$ sec per year.

10.10 Constant Acceleration

Although in special relativity we are restricted to inertial (that is, nonaccelerating) reference frames, and working with accelerations through the field equations of general relativity would require math far beyond the scope of this course, the case of constant acceleration is actually still within our reach.[64] The calculation is, however, a bit long and sinuous, involving several tangential observations, so we will tackle it in four distinct steps:

Step 1: Deriving a relativistic expression for velocity from our previous result for momentum.

Step 2: From this relativistic expression for velocity, deriving a relativistic expression for acceleration.

Step 3: Relating this relativistic acceleration to the acceleration familiar to us from Newtonian physics.

Step 4: Solving the resulting set of equations to determine the motion of the object experiencing the constant acceleration.

[64] For a more sophisticated and complete, but still concise, treatment of the case of constant acceleration, including the corresponding metric and gravitational field, you might check out Claude Semay, "Observer with a constant proper acceleration," `https://arxiv.org/abs/physics/0601179`.

Step 1: First, we need a relativistic expression for velocity. Recall from §10.7 and §10.7.1 that momentum, like all vector quantities in relativity, is a four-vector that has a temporal (p_t) as well as three spatial (p_x, p_y, p_z) components. If, as usual, we restrict ourselves to motion along the x axis, then $p_y = p_z = 0$ and we need concern ourselves only with

$$p_t = \gamma mc \qquad \text{and} \qquad p_x = \beta\gamma mc$$

Now, the Newtonian momentum was $p = mv$, so that velocity could be expressed as $v = p/m$, and we may similarly take p/m to define our relativistic velocity vector.[65] The conventional notation for this velocity four-vector is u:

$$u_t = \frac{p_t}{m} = \frac{\gamma mc}{m} = \gamma c \qquad\qquad u_x = \frac{p_x}{m} = \frac{\beta\gamma mc}{m} = \beta\gamma c$$

These relations for the components of u are correct, but not terribly illuminating; to re-express them in a way that makes more direct physical sense, note that

$$u_x = \beta\gamma c = \frac{v}{c}\,\gamma c = \gamma v = \gamma\frac{dx}{dt} \tag{10.102}$$

where dx is the spatial displacement that the moving object undergoes during the time interval dt from "our" perspective, that is, *according to an observer who sees the object moving at velocity* u. In the frame of the object itself, the displacement would of course be zero, and the time interval would be the proper time $d\tau$. And dt and $d\tau$ will be related by simple time dilation:[66]

$$d\tau = \frac{1}{\gamma}\,dt$$

We may therefore rewrite eq. (10.102) as

$$u_x = \gamma\frac{dx}{dt} = \frac{dx}{\frac{1}{\gamma}\,dt} = \frac{dx}{d\tau}$$

This expression for u_x is much easier to interpret physically: the relativistic velocity, like the Newtonian velocity, is a time derivative of location — the rate at which the object's location is changing — , except that the relativistic version involves a derivative specifically with respect to the proper time, the time on the object's own clock. The other difference is that the relativistic

[65] We had already reasoned out a relativistic expression for velocity in §10.7, but the reasoning we used then, when we were taking baby steps, was arguably a bit hokey. Our present derivation is more righteous.

[66] If this isn't clear, refer back to eq. (10.82) on p.565.

velocity also has a temporal component $u_t = \gamma c$, which we can similarly re-express as a derivative with respect to τ:

$$u_t = \gamma c = \gamma c \frac{dt}{dt} = c \frac{dt}{\frac{1}{\gamma} dt} = c \frac{dt}{d\tau}$$

Pulling this all together, and relaxing the restriction that the motion is confined to the x axis, we have as our result for the components of the relativistic four-velocity

$$u_t = c \frac{dt}{d\tau} \qquad u_x = \frac{dx}{d\tau} \qquad u_y = \frac{dy}{d\tau} \qquad u_z = \frac{dz}{d\tau} \tag{10.103}$$

In the nonrelativistic limit $v \ll c$, relativistic quantities (at least their the spatial components) should always reproduce the corresponding Newtonian quantities, and this is in fact true for our result for the relativistic velocity: when $v \ll c$, $\gamma \approx 1$, so that

$$d\tau = \frac{1}{\gamma} dt \approx dt$$

and

$$u_x = \frac{dx}{d\tau} \approx \frac{dx}{dt} \qquad u_y = \frac{dy}{d\tau} \approx \frac{dy}{dt} \qquad u_z = \frac{dz}{d\tau} \approx \frac{dz}{dt}$$

Note also that the usual combination[67]

$$(t \text{ component})^2 - \Big((x \text{ component})^2 + (y \text{ component})^2 + (z \text{ component})^2 \Big) \tag{10.104}$$

is invariant for our four-velocity (10.103):

$$\begin{aligned} u_t^2 - (u_x^2 + u_y^2 + u_z^2) &= \left(c \frac{dt}{d\tau} \right)^2 - \left(\left(\frac{dx}{d\tau} \right)^2 + \left(\frac{dy}{d\tau} \right)^2 \left(\frac{dz}{d\tau} \right)^2 \right) \\ &= \frac{c^2 \, dt^2 - (dx^2 + dy^2 + dz^2)}{d\tau^2} \end{aligned}$$

[67] When we have encountered this combination before, we have had a c^2 with the time term: $c^2 \, dt^2 - (dx^2 + dy^2 + dz^2)$ and the like. The presence or absence of the c^2 with the time term is, however, merely a matter of how the time component of the vector in question is defined: all of the components of a vector really should have the same physical dimensions, so the components of, for example, a position vector in spacetime should be, not (t, x, y, z), but (ct, x, y, z), so that all of the components have length dimensions. The c^2 in $c^2 \, dt^2$ is then built into the time component $c \, dt$.

which, by eq. (10.81) on p.564, reduces to

$$u_t^2 - (u_x^2 + u_y^2 + u_z^2) = \frac{c^2\, d\tau^2}{d\tau^2} = c^2 \tag{10.105}$$

In fact, the dot product of two four-vectors, say a and b, is given by the combination

$$a \cdot b = a_t b_t - (a_x b_x + a_y b_y + a_z b_z) \tag{10.106}$$

We leave it as an exercise for the reader to show that this dot product is invariant under Lorentz transforms [68] in the same way that the more familiar dot product

$$\mathbf{a} \cdot \mathbf{b} = a_x b_x + a_y b_y + a_z b_z$$

is invariant under rotations. The combination (10.104) is therefore simply the dot product of a four-vector with itself, that is, the squared magnitude of the four-vector. In particular, eq. (10.105) is telling us that the four-velocity u is a vector of constant magnitude.

Step 2: Since, by eqq. (10.103), the relativistic four-velocity u is the derivative of the spacetime location (ct, x, y, z) with respect to the proper time τ, we should obtain the relativistic four-acceleration simply by taking another derivative with respect to τ. Taking the acceleration a to be $du/d\tau$, we have

$$\begin{aligned} a_t &= \frac{du_t}{d\tau} = c\,\frac{d^2t}{d\tau^2} \\ a_x &= \frac{du_x}{d\tau} = \frac{d^2x}{d\tau^2} \\ a_y &= \frac{du_y}{d\tau} = \frac{d^2y}{d\tau^2} \\ a_z &= \frac{du_z}{d\tau} = \frac{d^2z}{d\tau^2} \end{aligned} \tag{10.107}$$

Recall that in Newtonian physics, the acceleration vector can always be split into two components, a component that lies along the direction of motion and corresponds to the rate at which the object is speeding up or slowing down (that is, to changes in the magnitude of the velocity), and a component perpendicular to the direction of motion that corresponds to the rate

[68] You can prove this fairly easily yourself, at least for our usual special case that the spatial components are restricted to the x axis (so that $a_y = a_z = b_y = b_z = 0$), by using

$$a'_t = \gamma(a_t - \beta a_x) \qquad b'_t = \gamma(b_t - \beta b_x)$$
$$a'_x = \gamma(-\beta a_t + a_x) \qquad b'_x = \gamma(-\beta b_t + b_x)$$

and working out $a' \cdot b'$.

at which the velocity is changing direction. One consequence of the four-velocity u having a constant magnitude is that it is, in the sense of the dot product (10.106), perpendicular to the four-acceleration a. To see this, we simply take the derivative of eq. (10.105) with respect to τ:

$$\begin{aligned}
\frac{d}{d\tau}\Big(u_t^2 - (u_x^2 + u_y^2 + u_z^2)\Big) &= \frac{d}{d\tau}\, c^2 \\
2u_t\frac{du_t}{d\tau} - \left(2u_x\frac{du_x}{d\tau} + 2u_y\frac{du_y}{d\tau} + 2u_z\frac{du_z}{d\tau}\right) &= 0 \\
2u_t a_t - (2u_x a_x + 2u_y a_y + 2u_z a_z) &= 0
\end{aligned}$$

so that

$$u_t a_t - (u_x a_x + u_y a_y + u_z a_z) = 0 \tag{10.108}$$

In Newtonian physics, where magnitudes involve the sums of squares like $a_x^2 + a_y^2 + a_z^2$, motion at a constant acceleration perpendicular to the velocity means that the velocity vector is changing only its direction, veering at a constant rate, with the result that the trajectory is circular. Without even having done out the calculation, we therefore expect that in relativity, where magnitudes involve the *differences* of squares of the form $u_t^2 - (u_x^2 + u_y^2 + u_z^2)$, a constant acceleration perpendicular to the velocity will result in a *hyperbolic* trajectory. In fact, as we will work out below, instead of getting circular motion where x and y are cos and sin as functions of time, we will find that x and t are hyperbolic functions — cosh and sinh — of the proper time τ.

Step 3: Now that we have a relativistic expression for acceleration, we need to relate it to the acceleration familiar to us from Newtonian physics. The first thing to get straight is what we mean by constant acceleration in relativity, and the sensible definition would be an acceleration of constant magnitude in the sense of eq. (10.104). That is, we want an acceleration such that

$$a_t^2 - (a_x^2 + a_y^2 + a_z^2) = \text{const} \tag{10.109}$$

Next, we need to determine the value of this constant in terms of the Newtonian acceleration, which, to avoid confusion with the relativistic four-acceleration a, we will denote by A. Since the combination on the left-hand side of eq. (10.109) is invariant, the constant on the right-hand side will have the same value in all reference frames; to determine that value, we need merely choose a frame in which the comparison to the Newtonian acceleration will be easy (or at least feasible), and the most likely candidate would seem to be the frame of the object itself: in the object's own frame, the object

is at rest, so that $\gamma = 1$ and $\tau = t$. In this frame, we therefore have

$$a_t = c\frac{d^2t}{d\tau^2} = c\frac{d^2t}{dt^2} = 0$$
$$a_x = \frac{d^2x}{d\tau^2} = \frac{d^2x}{dt^2} = A_x^2$$
$$a_y = \frac{d^2y}{d\tau^2} = \frac{d^2y}{dt^2} = A_y^2$$
$$a_z = \frac{d^2z}{d\tau^2} = \frac{d^2z}{dt^2} = A_z^2$$

so that eq. (10.109) becomes

$$a_t^2 - (a_x^2 + a_y^2 + a_z^2) = 0 - (A_x^2 + A_y^2 + A_z^2) = -A^2 \tag{10.110}$$

where A here represents the magnitude of the Newtonian acceleration in the usual three-dimensional spatial sense.

Step 4: Eqq. (10.105), (10.108), and (10.110) — which state that the velocity is of constant magnitude, that the acceleration is perpendicular to the velocity, and that the acceleration is of constant magnitude — reduce, if we restrict ourselves to motion along the x axis, to

$$u_t^2 - u_x^2 = c^2 \tag{10.111a}$$
$$u_t a_t - u_x a_x = 0 \tag{10.111b}$$
$$a_t^2 - a_x^2 = -A^2 \tag{10.111c}$$

where

$$u_t = c\frac{dt}{d\tau} \qquad a_t = c\frac{du_t}{d\tau} = c\frac{d^2t}{d\tau^2}$$
$$u_x = \frac{dx}{d\tau} \qquad a_x = \frac{du_x}{d\tau} = \frac{d^2x}{d\tau^2}$$

Our task is now to solve eqq. (10.111) to obtain x and t as functions of τ and thus, by eliminating τ, to determine the trajectory $x(t)$. Not every squirrel that sets out to cross the road makes it to the other side, of course, and roadkill might seem like the more likely outcome if you consider that eqq. (10.111) are not only coupled differential equations, but *nonlinear* coupled differential equations, for solving which there are no general techniques. But fortunately it turns out that eqq. (10.111) are not only tractable, but have relatively simple solutions.

What we want as a first step is to decouple the equations, so that we have an equation in terms of only x or only t components. To accomplish this, we can use eq. (10.111b) to eliminate a_t from eq. (10.111c). Using

$$a_t = \frac{u_x}{u_t}\,a_x$$

from eq. (10.111b) in eq. (10.111c) yields

$$\begin{aligned} -A^2 &= a_t^2 - a_x^2 \\ &= \left(\frac{u_x}{u_t}\, a_x\right)^2 - a_x^2 \\ &= \left(\left(\frac{u_x}{u_t}\right)^2 - 1\right) a_x^2 \\ &= \frac{u_x^2 - u_t^2}{u_t^2}\, a_x^2 \end{aligned}$$

Since this equation involves u_t as well as u_x and a_x, it might not seem that we've made all that much progress, but now we can use eq. (10.111a) to eliminate u_t. Substituting in $u_t^2 = c^2 + u_x^2$, we have

$$\begin{aligned} -A^2 &= \frac{u_x^2 - (c^2 + u_x^2)}{c^2 + u_x^2}\, a_x^2 \\ &= -\frac{c^2}{c^2 + u_x^2}\, a_x^2 \end{aligned}$$

which yields

$$a_x^2 = \frac{A^2}{c^2}(c^2 + u_x^2) = A^2\left(1 + \frac{u_x^2}{c^2}\right)$$

If we take our acceleration to be in the positive x direction, then $a_x > 0$, so that we want the positive root:

$$a_x = A\sqrt{1 + \frac{u_x^2}{c^2}}$$

We can now integrate this equation by using $a_x = du_x/d\tau$ and separating variables:

$$\begin{aligned} \frac{du_x}{d\tau} &= A\sqrt{1 + \frac{u_x^2}{c^2}} \\ \frac{du_x}{\sqrt{1 + u_x^2/c^2}} &= A\, d\tau \\ \int_0^{u_x} \frac{du_x}{\sqrt{1 + u_x^2/c^2}} &= \int_0^{\tau} A\, d\tau \end{aligned}$$

where we have, for simplicity, supposed that we have set up our axes so that the object is at rest ($u_x = 0$) at $\tau = 0$. The integration on the left-hand side

may well not be one you remember off the top of your head, but it turns out to be an inverse hyperbolic sine:

$$\int \frac{dz}{\sqrt{1+(z/c)^2}} = c\sinh^{-1}\frac{z}{c}$$

We thus have

$$\begin{aligned} c\sinh^{-1}\frac{u_x}{c}\Big|_0^{u_x} &= A\tau\Big|_0^{\tau} \\ c\sinh^{-1}\frac{u_x}{c} &= A\tau \\ u_x &= c\sinh\frac{A\tau}{c} \end{aligned} \tag{10.112}$$

Since $u_x = dx/d\tau$, we can integrate this one more time to obtain [69]

$$\begin{aligned} \frac{dx}{d\tau} &= c\sinh\frac{A\tau}{c} \\ \int_0^x dx &= \int_0^\tau d\tau\, c\sinh\frac{A\tau}{c} \\ x\Big|_{x_0}^{x} &= c\,\frac{c}{A}\cosh\frac{A\tau}{c}\Big|_0^{\tau} \\ x - x_0 &= \frac{c^2}{A}\left(\cosh\frac{A\tau}{c} - 1\right) \end{aligned} \tag{10.113}$$

where we have supposed that the object is at $x = x_0$ at $\tau = 0$.

We could go through a virtually identical routine to obtain u_t and t as functions of τ, but it is easier, now that we have the result (10.112) for u_x, to go back to eq. (10.111a) to obtain [70]

$$\begin{aligned} u_t^2 &= c^2 + u_x^2 \\ &= c^2 + \left(c\sinh\frac{A\tau}{c}\right)^2 \\ &= c^2\left(1 + \sinh^2\frac{A\tau}{c}\right) \\ &= c^2\cosh^2\frac{A\tau}{c} \end{aligned}$$

[69] Recall that the hyperbolic functions are free from the annoying signs that occur in the nonhyperbolic trig functions: while $\int \sin = -\cos$ and $\int \cos = \sin$, $\int \sinh = \cosh$ and $\int \cosh = \sinh$, and of course likewise for the derivatives.

[70] Recall that while $\sin^2 + \cos^2 = 1$, $\cosh^2 - \sinh^2 = 1$.

Since t will increase as τ increases, $u_t = c\,dt/d\tau > 0$, so that we want the positive root. Thus

$$u_t = c\cosh\frac{A\tau}{c} \tag{10.114}$$

$$c\frac{dt}{d\tau} = c\cosh\frac{A\tau}{c}$$

$$\int_0^t dt = \int_0^\tau d\tau\,\cosh\frac{A\tau}{c}$$

$$dt\Big|_0^t = \frac{c}{A}\,\sinh\frac{A\tau}{c}\Big|_0^\tau$$

$$t = \frac{c}{A}\,\sinh\frac{A\tau}{c} \tag{10.115}$$

where we have supposed that our time axis is set up so that $t = 0$ at $\tau = 0$.

To summarize, we have arrived at the solutions

$$u_x = c\sinh\frac{A\tau}{c} \tag{10.116a}$$

$$u_t = c\cosh\frac{A\tau}{c} \tag{10.116b}$$

$$x - x_0 = \frac{c^2}{A}\left(\cosh\frac{A\tau}{c} - 1\right) \tag{10.116c}$$

$$t = \frac{c}{A}\,\sinh\frac{A\tau}{c} \tag{10.116d}$$

First note that, as advertised, the trajectory is hyperbolic: eqq. (10.116d) and (10.116c) yield

$$\sinh\frac{A\tau}{c} = \frac{At}{c}$$

$$\cosh\frac{A\tau}{c} = 1 + \frac{A(x - x_0)}{c^2}$$

so that

$$\left(1 + \frac{A(x - x_0)}{c^2}\right)^2 - \left(\frac{At}{c}\right)^2 = \cosh^2\frac{A\tau}{c} - \sinh^2\frac{A\tau}{c} = 1 \tag{10.117}$$

which is indeed a hyperbola.

Second, note that although the object continually experiences a constant acceleration, it's velocity never exceeds the speed of light:

$$v = \frac{dx}{dt} = \frac{dx/d\tau}{dt/d\tau} = c\,\frac{dx/d\tau}{c\,dt/d\tau} = c\,\frac{u_x}{u_t} = c\,\frac{c\sinh A\tau/c}{c\cosh A\tau/c} = c\tanh\frac{A\tau}{c} \tag{10.118}$$

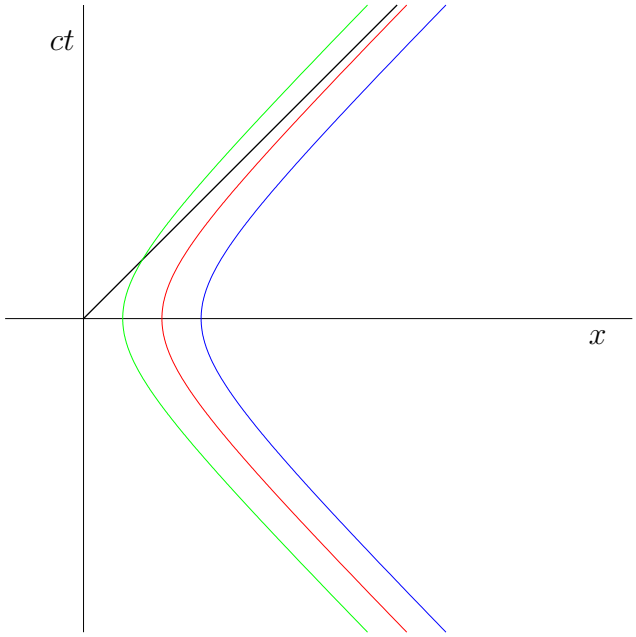


Figure 10.13: So Long, Sucker!

which asymptotically approaches $\pm c$ as τ (and hence t) goes to $\pm\infty$. The situation is in some ways similar to Zeno's paradox: although the object, in its own frame, is always experiencing a constant acceleration A — it would feel as though it were, equivalently, perpetually experiencing a constant gravitational field corresponding to a gravitational acceleration $g = A$ — , as it gets ever closer to the speed of light it experiences ever greater time dilation, so that the equal time intervals that are yielding equal increments in velocity from the object's perspective are, from our perspective, taking ever longer. Stated the other way around, in what are equal time intervals to us, the object's gains in velocity become ever smaller.[71]

Finally, you might naively expect that since the accelerating object never reaches the speed of light, a light ray would always, given enough time, be able to catch up with it. This turns out, however, not to be true: if the object has enough of a head start, the light ray will never catch up to it. Fig. (10.13) is a spacetime diagram of the trajectories of a light ray and of three objects experiencing the same constant acceleration but starting from different initial locations x_0. The light ray moves in the positive x direction starting from $x = 0$ at $t = 0$, so that its trajectory is $x = ct$ — the black line

[71] This is effect is also not unlike the difference in times for an object to fall into a Schwarzschild black hole: on the clock of the falling object, the time to reach the event horizon is finite, but to an observer outside the black hole it is infinite — the object never reaches the black hole from that observer's perspective.

of unit slope. The green, red, and blue curves are the hyperbolic trajectories of eq. (10.117) for, respectively, the cases

$$x_0 = \frac{1}{2}\frac{c^2}{A},\ \frac{c^2}{A},\ \text{and}\ \frac{3}{2}\frac{c^2}{A}$$

In each of these three cases the object starts from rest at x_0 at $t = 0$, that is, with a head start of x_0 on the light ray.[72] The intersection of black and green trajectories is the point where the light ray catches up with the object that had a head start of only $c^2/2A$, but the light ray never catches up with the objects moving along the red and blue trajectories. The critical case occurs when the asymptote of the hyperbolic trajectory is the trajectory $x = ct$ of the light ray. From eq. (10.117), the asymptote is given by

$$\begin{aligned}
1 + \frac{A(x - x_0)}{c^2} &= \frac{At}{c} \\
\frac{A(x - x_0)}{c^2} &= \frac{At}{c} - 1 \\
x - x_0 &= \frac{c^2}{A}\left(\frac{At}{c} - 1\right) \\
&= ct - \frac{c^2}{A} \\
x &= ct + \left(x_0 - \frac{c^2}{A}\right)
\end{aligned}$$

So if the object starts from rest with a head start of $x_0 \geq c^2/A$, the light ray will never catch up to it. If you plug the numbers in, for $A = g$ (the acceleration due to gravity on Earth), the critical head start $x_0 = c^2/A$ works out to about one light-year (a light-year being $c \times 1\,\text{yr}$, the distance light travels in one year).

[72] To make their hyperbolic aspect more apparent, in fig. (10.13) we have shown the trajectories of the three objects for times $t < 0$ as well, so more precisely the x_0 are the turn-around points where the objects have ceased their backwards motion, come to rest, and then begin moving forward. But you knew what we meant.

10.11 Problems

An asterisk indicates a numerical problem that will require a calculator; in other numerical problems the numbers are nice enough that you can do them by hand, as long as you remember the Pythagorean triplets

$$3^2+4^2=5^2 \qquad 7^2+24^2=25^2 \qquad 49^2+1200^2=1201^2$$
$$5^2+12^2=13^2 \qquad 9^2+40^2=41^2$$

*1. A light-year is a unit of *length*: it is the distance that light travels in one year — 2.99792458×10^8 m/s $\times$ 1 yr, which, if you do the conversions, works out to about 9.46×10^{15} m. In the context of relativity, it is often easier to work with distances in light-years (ly) than in meters, because velocities are often known as fractions of the speed of light, and 1 ly is literally 1 yr times the speed of light. For example, if you travel a distance of 3 ly at half the speed of light, the time for this trip will be

$$\text{time}=\frac{\text{distance}}{\text{velocity}}=\frac{3\text{ ly}}{\frac{1}{2}c}=6\text{ yr}$$

because

$$\frac{3\text{ ly}}{\frac{1}{2}c}=\frac{3\text{ yr}\times c}{\frac{1}{2}c}$$

and the c's therefore cancel out. You should get used to working with light-years.

Suppose the Enterprise makes a trip from Earth to a star system 36.0 ly away. Determine

i. The time the trip takes according to an observer on Earth

ii. The time the trip takes according to an observer on the Enterprise

iii. The distance traveled on the way to the star system according to an observer on the Enterprise

if the Enterprise makes the trip at

(a) 1% of the speed of light.

(b) 10% of the speed of light.

(c) 70% of the speed of light.

(d) 99% of the speed of light.

(e) 99.99% of the speed of light.

(f) 99.9999% of the speed of light.

2. You decide to brush back a batter by throwing a bean-ball at $\frac{3}{5}c$.

   (a) According to the batter, what is the ratio of the volume of the softball to its rest volume?

   (b) According to you, what is the ratio of the volume of the softball to its rest volume?

   (c) If the softball were sentient, how would it perceive its volume in flight to differ from its rest volume?

   (d) How long will you assert that it takes the ball to reach the batter's head if a measuring tape laid out along the field shows the distance from the rubber to home plate to be ℓ?

   (e) How long will the batter assert that it takes the ball to reach his or her head?

   (f) How long would the ball assert that it took to reach the batter's head?

3. Back in elementary school, you used to express a romantic interest in someone by throwing rocks at him or her on the playground during recess. And then all your friends would help things along by singing, "X and Y, sittin' in a tree, K-I-S-S-I-N-G! . . . " Life was so much simpler in those days. Although the whole tree thing was a little confusing. Anyway, suppose that to express a deep affection you heave a big, pointy rock at the object of your desires. According to the playground monitor, who, like all playground monitors, stays completely at rest in a quiet corner of the playground throughout the recess, the rock takes $\frac{13}{100}$ sec to reach the target. According to the rock itself, it takes $\frac{1}{20}$ sec to reach the target.

   (a) What is the speed of the rock according to the playground monitor?

   (b) How far does the rock travel on its way to the target according to the playground monitor?

   (c) How far does the rock perceive itself to travel on its way to the target?

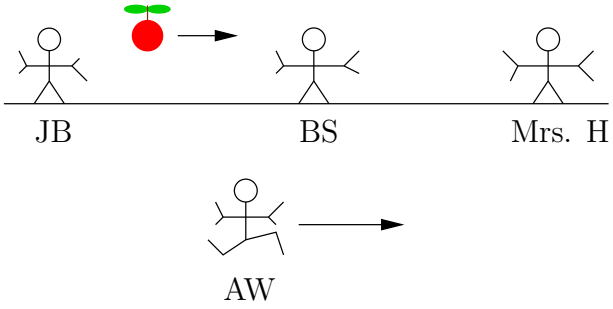


Figure 10.14: Problem 4

4. (Based on a true story.) JB[73] heaves an apple at BS, as shown in fig. (10.14). The apple completely misses BS and, 4.0 sec and 4.0×10^8 m later (according to JB), pegs Mrs. H. in the back of the head.

   (a) i. If AW is observing all this as he runs by at $\frac{3}{5}c$, what are the apple's time of flight and displacement according to AW?

   ii. How fast is the apple moving according to JB?

   iii. How fast is the apple moving according to AW?

   iv. In light of your answers to #4(a)ii and #4(a)iii, make physical sense of the relation between your results for the apple's time of flight and displacement according to JB and according to AW

[73] This of course refers to *the* JB — or rather, the *other* JB — , that true southern gentleman with a heart of gold but a propensity for mischief that left many convinced he was born to the gallows. Tom Sawyer and Huck Finn had nothing on our hero; the infractions cited in the disciplinary file he amassed during his four years with us are unlikely ever to be surpassed in number, color, or variety, and a suitably epic description of them would be far beyond not only our muse but very likely even Mr. Twain's. In addition to such pedestrian fare as food-fights and moonings, including at least one drive-by mooning from the window of a bus, JB displayed his prowess in countless original and exotic ways, from the impressive feat of denting the door of a van with a water balloon fired from a water winger to the artful enterprise of the remotely activated electric cattle prod under the cushions of his dorm-room sofa.

We cannot, however, allow the mention of his name to pass without relating one particular story. When JB was in Lower School, he broke his leg snowboarding — badly enough that it had to be set with pins and secured in a full-length cast from foot to hip. It was not many weeks after this that we were surprised one evening to see him striding down the halls of Lower without the cast.

"You got your cast off!"

"Yeah," he replied, with possibly the widest smile we have ever seen.

"Great! What did the doctor say?"

"Doctor? What doctor?"

And so it came to light that JB, simply tired of having to hobble around with the cast on, had spent some four hours in his room that afternoon whittling it away with his pocket knife. This hard-won liberty was, however, short-lived: the responsible authorities saw to it that he was back in a new cast the next day.

Anyway, the incident to which the problem refers would probably remained lost to

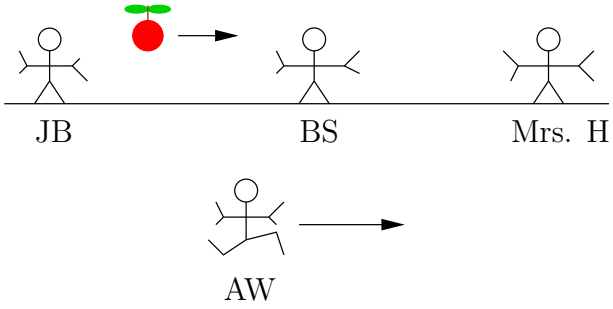


Figure 10.15: Problem 4

(b) Suppose that instead the 4.0 sec and 4.0×10^8 m were according to AW, not JB. What would be the apple's time of flight and displacement according to JB?

(c) Okay, back to the 4.0 sec and 4.0×10^8 m being the time and displacement according to JB.

i. Is the separation between the throwing of the apple and its hitting Mrs. H. time-like, space-like, or light-like? (If we haven't covered the time-like–space-like–light-like business yet, just go on to the following parts and we'll deal with this part in class.)

ii. Is there a frame in which these two events happen at the same place? If not, why not? If so, how fast and in what direction is this frame moving relative to JB?

iii. Is there a frame in which these two events happen at the same time? If not, why not? If so, how fast and in what direction is this frame moving relative to JB?

history had we not had the good fortune to have had JB and AW in the same class some years ago. When we were working through a projectile-motion problem, AW interrupted and regaled us all with the story of how a certain someone once threw an apple at BS, but that certain someone overshot, with the result that the apple thrown by that certain someone pegged Mrs. H. on the back of the head. AW had no sooner finished the story than JB, with his characteristic impulsiveness, and with all of the uncalculating innocence of the day he was born, blurted out, "I thought I hit her on the leg?" The entire class turned in unison and stared at him in shocked disbelief — not, of course, at the enormity of his having struck Mrs. H., but at his having so recklessly and gratuitously revealed himself as the perpetrator. A long pause followed, during which JB ruefully reflected on the grotesque imprudence of his spontaneous confession, nervously fidgeting under the continuing stare of the class, until finally he broke the silence by sheepishly inquiring, "Um, you're not going to tell Mr. H., are you?"

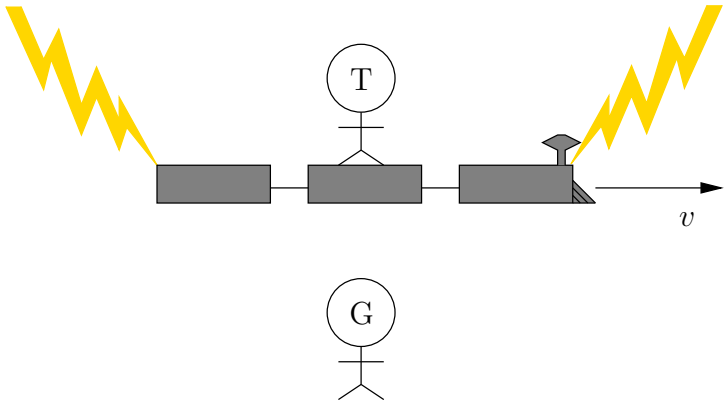


Figure 10.16: Problem 5

5. A classic thought experiment to demonstrate the relativity of simultaneity is the old lightning-bolts-and-train problem: lightning bolts strike the front and back of a train of rest length ℓ, as shown in fig. (10.16). To an observer T on the train (which is moving at velocity v), the strikes occur simultaneously.

   Show that the strikes are not simultaneous according to an observer G standing at rest on the ground. Also determine which strike occurs first according to this observer.

6. (The twin paradox.[74]) Dork A and dork B are coæval.[75] Dork A remains on Earth while dork B travels out to a star 82 ly away at a speed of $\frac{40}{41}c$, abruptly turns around, and then returns at the same $\frac{40}{41}c$.

   (a) According to dork A, how long does the trip take, and how far has dork B traveled?

   (b) According to dork B, how long does the trip take, and how far has dork B traveled?

   (c) How does this show that a lot of what you see on *Star Trek* and similar shows is malarkey?

   (d) Why can't we equally well look at things from dork B's perspective and argue that dork A is the one who's moving and therefore the one for whom the time is shorter?

[74] For more on the twin paradox, specifically a calculation of the effects of the acceleration necessary to reverse direction, you might look at Lorenzo Iorio, *An Analytic Treatment of the Clock Paradox in the Framework of the Special and General Theories of Relativity*, `https://arXiv.org/abs/physics/0405038`. Or simply go to `https://arxiv.org` and search for papers about the twin paradox: even a hundred years later, there is still a steady stream of papers being written about it.

[75] The dictionary is your friend. Be prepared, however, for some disappointment; this word sounds much cooler than it actually is.

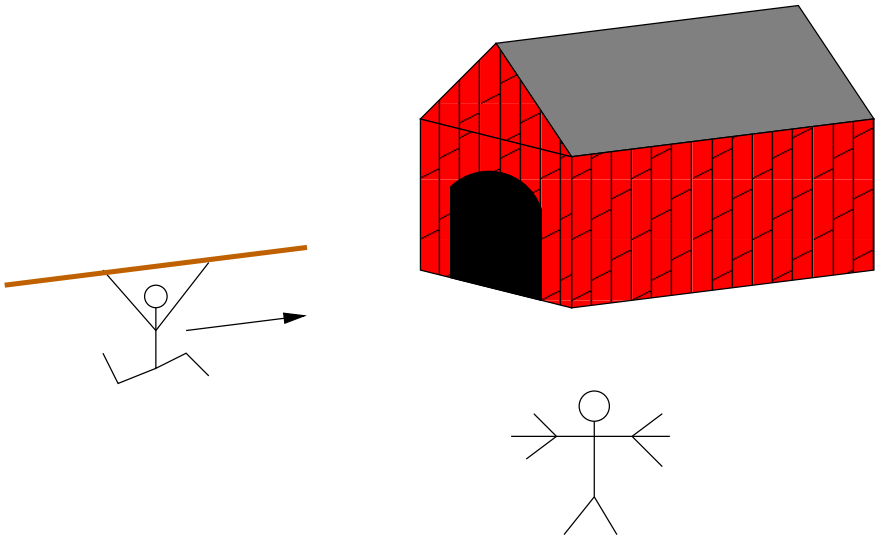


Figure 10.17: Problem 7

7. (The classic pole-in-barn paradox.) A pole is equal in length to a barn when both are at rest. One dork picks up the pole and, holding it horizontally along the direction of motion, runs into, through, and out the back of the barn, as shown with great artistry in fig. (10.17). The dork running with the pole argues that since the barn was contracted in his frame of reference, there was no time at which the entire length of the pole could have been inside the barn. Another dork, standing outside the barn, argues that it was the pole that was moving and therefore contracted, so that there was a time when the pole was entirely within the barn. Who is right? Will violence be necessary? Is there some way to set up cameras to resolve the issue? What if you took time-stamped pictures with synchronized cameras mounted on the front and back of the barn? And used these cameras to take pictures of synchronized clocks mounted on the ends of the pole?

8. According to a traumatized observer standing on the ground, a newly emerged cicada flies at a distance ℓ in a straight line at a constant speed in time t before being squashed on the windshield of a tractor-trailer. How long is the flight of the cicada on its own clock?

9. (The occasion of a Darwin Award — but only an Honorable Mention, since the candidate only lost an eye.) After a few beers, some rednecks decide it would be fun to do the William-Tell thing by shooting beer cans off of each other's heads with a bow and arrow. To a cynically amused spectator standing at rest on the ground, the arrow seems to be traveling at speed V and to take time T to pass through the beer can, which for simplicity we will treat as a point. What is the arrow's rest length?

10. You are in a hurry and want to be able to make it to a star system a distance ℓ away in what will, by your clock on the spaceship, seem like time T.

    (a) Determine the fraction of the speed of light you must travel.

    (b) How much time will elapse during your trip according to people on Earth?

    (c) Show that there is no limit on how short you can make the trip time T and determine what happens to the time elapsed according to people on Earth as $T \to 0$.

11. For this problem we hark back to classical antiquity and Æsop's fable of the tortoise and the hare.[76] As you may recall, the fox was officiating at a race between the tortoise, who ran at a steady pace right from the starting gun, and the hare, who first took a nap before dashing off at a greater but likewise constant speed. We will take the fox to be a stationary observer, from whose perspective the tortoise covers the 60 ly race course at a steady $\frac{5}{13}c$ and the hare, after napping, dashes off at a constant $\frac{12}{13}c$. The fox, the tortoise, and the hare have all synchronized their wristwatches before the race.

    (a) For how long, on the fox's watch, can the hare nap if the tortoise and hare are to reach the finish line in a tie according to the fox?

    (b) For how long, on the hare's watch, can the hare nap if the tortoise and hare are to reach the finish line in a tie according to the fox?

    (c) For how long, on the fox's watch, can the hare nap if the tortoise and hare are to reach the finish line in a tie according to their own wristwatches? That is, if the time on the tortoise's watch when the tortoise crosses the finish line is to equal the time on the hare's watch when the hare crosses the finish line?

    In this case, who wins the race according to the fox? According to the tortoise?

    (d) If the hare can instead run arbitrarily close to the speed of light, for how long can the hare nap if the tortoise and hare are to reach the finish line in a tie according to the fox?

    (e) If the hare can run arbitrarily close to the speed of light, for how long can the hare nap if the tortoise and hare are to reach the finish line in a tie according to their own wristwatches?

[76] This problem suggested by *Two New Variations on the Twin Pseudoparadox*, Jean Marc Levy Leblond, `https://arxiv.org/abs/1504.06186`.

12. One day you decide "**** this!" and get on a rocket ship and fly straight away from Earth at a steady $\frac{3}{5}c$. After you have been gone a year on your clock, you are reminiscing about the friends you left behind and decide to send them a picture of you — as you moon the camera. When the picture arrives by radio signal back on Earth, how much time, according to the people on Earth, has elapsed since you left the Earth?

13. Event B occurs at a distance ℓ farther down the positive x axis, and at a time t after, event A.

    (a) For what values of ℓ and t (that is, for what relationship between ℓ and t) will the separation between these two events be light-like?

    (b) i. For what values of ℓ and t will the separation between these two events be time-like?

    ii. Assuming that the separation is time-like, determine the velocity of the reference frame in which the two events occur at the same place.

    (c) i. For what values of ℓ and t will the separation between these two events be space-like?

    ii. Assuming that the separation is space-like, determine the velocity of the reference frame in which the two events occur at the same time.

14. In this problem we will restrict ourselves to frames F and F' in relative motion along an x axis and to events that occur on that axis.

    (a) Determine whether each of the following cases can occur. If it can, determine the conditions under which it is possible and the relation or relations that must hold between F and F'. Either way, prove your assertion and rationalize it physically.

        i. If event A occurs before event B in frame F, can it occur at the same time as B in frame F'?

        ii. If event A occurs before event B in frame F, can it occur after B in frame F'? If so, does this violate causality? That is, does it allow a future event to influence a past event?

        iii. If event A occurs farther down the x axis than event B (that is, $x_A > x_B$) in frame F, can A and B occur at the same place in frame F'?

        iv. If event A occurs farther down the x axis than event B (that is, $x_A > x_B$) in frame F, can B occur farther down the x axis than A in frame F'?

    (b) Show that if the time interval and spatial displacement between two events in frame F are both the same as in frame F', then F and F' are the same frame.

    (c) Show that if any two of Δx, Δt, $\Delta x'$, and $\Delta t'$ vanish, then either the frames F and F' are the same or all of Δx, Δt, $\Delta x'$, and $\Delta t'$ vanish.

15. (a) By doing a Taylor expansion of

$$\gamma = \sqrt{\frac{1}{1-\beta^2}}$$

show that

$$\gamma \approx 1 + \tfrac{1}{2}\beta^2 \qquad \text{for} \qquad \beta = \frac{v}{c} \ll 1$$

and hence that

$$\frac{1}{\gamma} \approx 1 - \tfrac{1}{2}\beta^2 \qquad \text{for} \qquad \beta = \frac{v}{c} \ll 1$$

    (b) Show that for velocities very close to the speed of light (that is, when $1 - \beta \ll 1$)

$$\gamma \approx \sqrt{\frac{1}{2(1-\beta)}}$$

*16. (You may find the results of #15 helpful for this problem.) You set out to prove to a friend skeptical of the benefits of exercise that running does in fact increase your longevity. One hot and humid summer day, while your friend takes a snooze in the shade under a tree, you run around the tree in the blazing sun at a steady 12 mph = 5.4 m/s for what is, according to your wristwatch, one hour.

(a) For how long are you running according to your friend?

(b) How far do you perceive yourself to have run?

(c) How far have you run according to your friend?

(d) How much time will you have added to your life because of time dilation?

*17. An eV (electron volt) is a unit of energy: $1\,\text{eV} = 1.60217653 \times 10^{-19}\,\text{J}$. An atomic mass unit (u) is defined to be one twelfth of the mass of ^{12}C (carbon-12): $1\,\text{u} = 1.66053886 \times 10^{-27}\,\text{kg}$. Also, to save you the trouble of looking it up (even though you could probably use the exercise), the speed of light is $c = 2.99792458 \times 10^{8}\,\text{m/s}$.

(a) Using $E = mc^2$, determine the energy that is released when two deuterium nuclei (each 2.01410178 u) are fused into a helium-4 nucleus (4.00260325 u). Express your answer in both MeV and Joules.

(b) How many Joules of energy would therefore be released by the fusion of 1 kg of deuterium?

*18. At particle colliders, subatomic particles of relatively low mass are accelerated to very nearly the speed of light, so that their combined energy when smashed together head-on is sufficient to produce more exotic particles of much greater mass.

Suppose you could do this sort of thing on a macroscopic scale. How fast would you have to smash two 1 kg chickens together to make a 400 kg cow? Express your answer as a result for $1 - \beta$.

19. On p.568 we arrived at the relation for addition of velocities:

$$u' = \frac{u - v}{1 - uv/c^2}$$

where u and u' are the velocities of the object in frames F and F', respectively, and v is the velocity of F' relative to F.

Invert this relation to obtain a result for u in terms of u' and v. You should be able to just write down the result without doing any work.

20. The batter you beaned in #2 charges the mound at $\frac{1}{2}c$ and, while charging, throws the bat at you. If the bat is moving at $\frac{3}{4}c$ according to you, how fast is it moving according to the batter?

21. You put the pedal to the metal, peel out, and tear across the student parking lot at $\frac{2}{3}c$. As you are moving at this speed, you hurl a lacrosse ball at an innocent bystander.[77] Your throwing speed when you are at rest on the ground, and therefore the speed at which *you* see the ball moving, is $\frac{1}{2}c$. How fast is the ball approaching according to the bystander?

22. A starship under Itchy's command pursues Scratchy, who has just stolen vital plans for a flux capacitor. At one point in the ensuing dog fight, the two fly their ships directly at each other. According to you, who are observing everything from the Earth, Scratchy is flying at $\frac{1}{2}c$. According to Scratchy, Itchy is approaching at $\frac{3}{4}c$.

    (a) How fast is Itchy's ship moving according to you?

    (b) Itchy fires a volley of abuse at Scratchy. If the abuse has a muzzle velocity of $\frac{4}{5}c$, at what speed, according to you, does it close on Scratchy?

    (c) In return, Scratchy fires a photon torpedo at Itchy. If the torpedo is moving at the speed of light according to you, how fast is it moving according to Itchy?

23. On p.568 we arrived at the relation for addition of velocities:

$$u' = \frac{u - v}{1 - uv/c^2}$$

    where u and u' are the velocities of the object in frames F and F', respectively, and v is the velocity of F' relative to F.

    Show that if u and v do not exceed the speed of light, then neither does u'. See the footnote if you need a hint.[78]

24. Recall that in #4 you found that the apple's time of flight and displacement according to JB and according to AW were symmetric. Now use the relation for addition of velocities to make sense of this symmetry in terms of the velocity of the apple according to JB, the velocity of the apple according to AW, and the velocity of AW according to JB.

[77] Bystanders are notoriously innocent. One wonders whether there is any other kind of bystander.

[78] Since u' depends on the two independent variables u and v, a maxima-minima calculation will yield two simultaneous conditions for extremizing u' (*simultaneous* here in the algebraic sense of *simultaneous equations*).

25. One of the spectral lines of hydrogen has a wavelength of about 21 cm in the lab. If this spectral line is red-shifted to 42 cm in the light from a distant star, how fast is this star receding from us?

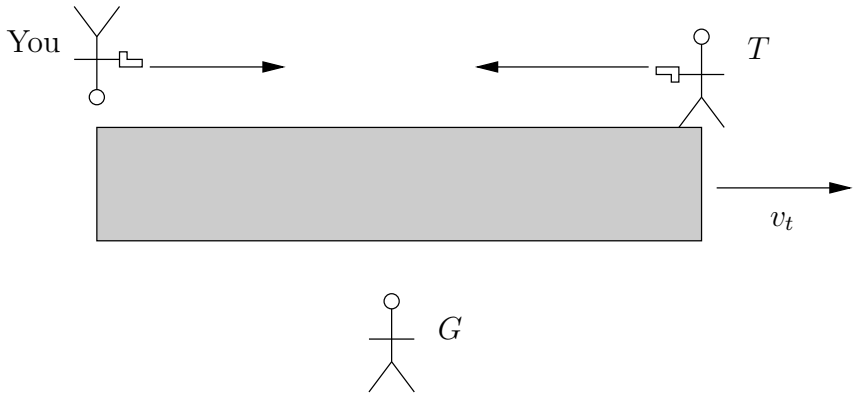


Figure 10.18: Problem 26

26. You step through the wrong door and find yourself in The Matrix. Specifically, as shown in fig. (10.18), you find yourself battling with a Smith (T) on top of the trailer of a moving 18-wheeler as another Smith (G) stands on the road watching. From the perspective of agent G, the 18-wheeler has length ℓ and is moving down the highway at speed v_t. You and agent T fire at each other with guns that both have a muzzle velocity v_b (that is, the speed of each bullet is v_b in the gun's own reference frame). From the perspective of G, you and T fire simultaneously. The bullets subsequently collide dead-on in midair.

    (a) From the perspective of G, how much time elapses between your firing and the collision? (This part might seem complicated, even downright ornery, but it's not bad as long as you are logical and methodical.)

    (b) From your perspective, did you and T fire at the same time? If so, explain why. If not, determine who fired first and the time between firings.

    (c) From your perspective, how much time elapses between your firing and the collision? (Your result will be fairly messy, but the calculation by which you arrive at that result will be quite tractable as long as you remain calm.)

27. Recall from §10.7 that the components of the relativistic four-vector velocity were, in terms of the proper time τ,

$$\left(\frac{c\,dt}{d\tau}, \frac{dx}{d\tau}, \frac{dy}{d\tau}, \frac{dz}{d\tau}\right)$$

Show that we therefore always have

$$u_t^2 - u_x^2 - u_y^2 - u_z^2 = c^2$$

28. Suppose that in the lab frame a mass m traveling at speed v_0 collides, completely inelastically, with an equal stationary mass m.

In the following parts, we will use the notation

$$\beta_0 = \frac{v_0}{c} \qquad\qquad \gamma_0 = \frac{1}{\sqrt{1-\beta_0^2}}$$

and likewise, for the center-of-mass velocity v_{cm},

$$\beta_{cm} = \frac{v_{cm}}{c} \qquad\qquad \gamma_{cm} = \frac{1}{\sqrt{1-\beta_{cm}^2}}$$

(a) Write down expressions for the spatial and temporal components p_x and p_t of the momentum of each mass in the lab frame.

(b) By means of the addition-of-velocities formula, relate the velocity of each mass in the center-of-mass frame to its velocity in the lab frame in terms of the (as yet unknown) velocity v_{cm} of the center of mass.

(c) Recall that the center-of-mass frame could more precisely be called the center-of-momentum frame — the frame in which the momenta are equal and opposite — which, when the masses are equal, means that the velocities are equal and opposite. By imposing the condition that your velocities in the center-of-mass frame are equal and opposite, show that the velocity v_{cm} of the center-of-mass frame is given by the following equivalent expressions:

$$\beta_{cm} = \frac{1-\sqrt{1-\beta_0^2}}{\beta_0} = \frac{\beta_0}{1+\sqrt{1-\beta_0^2}} = \frac{\gamma_0\beta_0}{\gamma_0+1}$$

and that thus

$$\gamma_{cm} = \sqrt{\tfrac{1}{2}(\gamma_0+1)}$$

(d) Show that in the nonrelativistic limit $v_0 \ll c$ this result for the velocity of the center of mass reduces to the expected Newtonian result.

(e) By applying the Lorentz transform (10.94),

$$p'_x = \gamma(p_x - \beta p_t) \qquad\qquad p'_t = \gamma(-\beta p_x + p_t)$$

to the momentum four-vector of each mass, show that in the center-of-mass frame we have

$$p'_x = \gamma_{cm}\gamma_0 m(v_0 - v_{cm}) \qquad\qquad E' = \gamma_{cm}\gamma_0 mc^2(1 - \beta_0\beta_{cm})$$

for the mass moving at v_0 in the lab frame and

$$p'_x = -\gamma_{cm} m v_{cm} \qquad\qquad E' = \gamma_{cm} mc^2$$

for the mass that is stationary in the lab frame.

(f) Show that imposing the condition that your results of #28e for spatial components of the momenta of the masses are equal and opposite in the center-of-mass frame yields the same result as #28c for v_{cm}.

(g) Show that the total energy in the center-of-mass frame is

$$mc^2\sqrt{2(\gamma_0 + 1)}$$

and that the combined mass after the collision is thus

$$m\sqrt{2(\gamma_0 + 1)}$$

29. There are formalisms for relativity that use an imaginary time coordinate. These formalisms, while in many respects cumbersome, have the advantage that spacetime becomes Euclidean: if, in place of the time coordinate t, we use $\tilde{t} = it$, where $i = \sqrt{-1}$, then $t = -i\tilde{t}$ and the invariant spatial interval between events (see §10.5)

$$\Delta x^2 - c^2 \Delta t^2$$

becomes

$$\Delta x^2 - c^2(-i\Delta\tilde{t})^2 = \Delta x^2 + c^2 \Delta \tilde{t}^2$$

which now looks just like the sum of squares that occurs in the Pythagorean theorem.

In this problem we will see that when an imaginary time coordinate is used the Lorentz transform can be regarded as a rotation by an imaginary angle in the xt plane. In other words, going from one reference frame to another is like rotating your perspective in the xt plane.

(a) Using the Euler relation

$$e^{i\theta} = \cos\theta + i \sin\theta$$

and the definitions of the hyperbolic functions

$$\begin{aligned} \sinh u &= \tfrac{1}{2}(e^u - e^{-u}) \\ \cosh u &= \tfrac{1}{2}(e^u + e^{-u}) \\ \tanh u &= \frac{\sinh u}{\cosh u} = \frac{e^u - e^{-u}}{e^u + e^{-u}} \end{aligned}$$

show that

$$\sinh iu = i \sin u \qquad \cosh iu = \cos u$$

and thus

$$\sin i\theta = i \sinh\theta \qquad \cos i\theta = \cosh\theta$$

(b) Recall that if the coordinate axes x', y' are rotated (in the ordinary spatial sense of a rotation) by a counterclockwise angle θ relative to the axes x, y, then

$$\Delta x' = \cos\theta\ \Delta x + \sin\theta\ \Delta y$$
$$\Delta y' = -\sin\theta\ \Delta x + \cos\theta\ \Delta y$$

Show that if we replace the y coordinate by ct and do a rotation by an imaginary angle $i\varphi$ instead of the real angle θ, these relations become

$$\Delta x' = \cosh\varphi\ \Delta x + i\sinh\varphi\ c\Delta t$$
$$c\Delta t' = -i\sinh\varphi\ \Delta x + \cosh\varphi\ c\Delta t$$

and hence, when expressed in terms of the imaginary time coordinate $\tilde{t} = it$,

$$\Delta x' = \cosh\varphi\ \Delta x + \sinh\varphi\ c\Delta\tilde{t}$$
$$c\Delta\tilde{t}' = \sinh\varphi\ \Delta x + \cosh\varphi\ c\Delta\tilde{t}$$

(c) Show that

$$\cosh\varphi = \frac{1}{\sqrt{1-\tanh^2\varphi}}$$

and that if we define

$$\beta = -\tanh\varphi$$
$$\gamma = \cosh\varphi = \frac{1}{\sqrt{1-\beta^2}}$$

the $x', \tilde{t}'$ transform can therefore be written as

$$\Delta x' = \gamma(\Delta x - \beta c\Delta\tilde{t})$$
$$c\Delta\tilde{t}' = \gamma(-\beta\Delta x + c\Delta\tilde{t})$$

30. This exercise will be an introductory adventure in the rapidity φ, an alternative measure of how fast an object is moving that, for $\beta = v/c$, is defined by

$$\tanh\varphi = \beta$$

(a) If you are not already familiar with these relations between the circular and hyperbolic functions, using the Euler relation

$$e^{i\theta} = \cos\theta + i\sin\theta$$

and the definitions of the hyperbolic functions

$$\begin{aligned}\sinh u &= \tfrac{1}{2}(e^u - e^{-u})\\ \cosh u &= \tfrac{1}{2}(e^u + e^{-u})\\ \tanh u &= \frac{\sinh u}{\cosh u} = \frac{e^u - e^{-u}}{e^u + e^{-u}}\end{aligned}$$

show that

$$\sinh iu = i\sin u \qquad \cosh iu = \cos u \qquad \tanh iu = i\tan u$$

and thus

$$\sin i\theta = i\sinh\theta \qquad \cos i\theta = \cosh\theta \qquad \tan i\theta = i\tanh\theta$$

(b) If you are not already familiar with this relation, show that while

$$\cos^2\theta + \sin^2\theta = 1$$

the corresponding relation among hyperbolic functions is[79]

$$\cosh^2 u - \sinh^2 u = 1$$

(c) If you are not already familiar with these relations, using

$$\begin{aligned}\sin(\theta_1 + \theta_2) &= \sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2\\ \cos(\theta_1 + \theta_2) &= \cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2\end{aligned}$$

show that

$$\begin{aligned}\sinh(u_1 + u_2) &= \sinh u_1\cosh u_2 + \cosh u_1\sinh u_2\\ \cosh(u_1 + u_2) &= \cosh u_1\cosh u_2 + \sinh u_1\sinh u_2\end{aligned}$$

[79] The relation $\cos^2\theta + \sin^2\theta = 1$, corresponding to $x^2 + y^2 = 1$ for the unit circle, is in fact why sin, cos, and their kindred are called the *circular functions*; the relation $\cosh^2 u - \sinh^2 u = 1$, corresponding to the hyperbola $x^2 - y^2 = 1$, is why sin, cos, and their kindred are called the *hyperbolic functions*.

(d) If you are not already familiar with these relations, show that

$$\tan(\theta_1 + \theta_2) = \frac{\tan\theta_1 + \tan\theta_2}{1 - \tan\theta_1 \tan\theta_2}$$
$$\tanh(u_1 + u_2) = \frac{\tanh u_1 + \tanh u_2}{1 + \tanh u_1 \tanh u_2}$$

If you need a hint, see the footnote.[80]

(e) Plot $\sinh u$, $\cosh u$, and $\tanh u$ — which you should be able to do without a calculator. Note in particular how $\tanh\varphi = \beta$ maps $-1 < \beta < 1$ onto $-\infty < \varphi < \infty$.

(f) At long last, the physics parts of the exercise: show that if the relation for relativistic addition of velocities on p.568,

$$u' = \frac{u - v}{1 - uv/c^2}$$

is instead written in terms of the corresponding rapidities

$$\varphi'_u = \tanh^{-1}\frac{u'}{c} \qquad \varphi_u = \tanh^{-1}\frac{u}{c} \qquad \varphi_v = \tanh^{-1}\frac{v}{c}$$

then it may be expressed simply as

$$\varphi'_u = \varphi_u - \varphi_v$$

That is, while velocities are not arithmetically additive in relativity, rapidities are.

(g) Show that in terms of the rapidity

$$\varphi = \tanh^{-1}\beta$$

the Lorentz transform

$$\Delta x' = \gamma(\Delta x - \beta c\Delta t)$$
$$c\Delta t' = \gamma(-\beta\Delta x + c\Delta t)$$

can be written

$$\Delta x' = \cosh\varphi(\Delta x - \tanh\varphi\, c\Delta t)$$
$$c\Delta t' = \cosh\varphi(-\tanh\varphi\, \Delta x + c\Delta t)$$

[80] Remember that

$$\tan\theta = \frac{\sin\theta}{\cos\theta}$$

And for the hyperbolic tangent, note that

$$\tanh i(\theta_1 + \theta_2) = i\tan(\theta_1 + \theta_2)$$

(h) Show that if one first Lorentz transforms from frame F to frame F' moving at velocity v' relative to F, and then Lorentz transforms from frame F' to frame F'' moving at velocity v'' relative to F', then in terms of the rapidities

$$\varphi' = \tanh^{-1}\frac{v'}{c} \qquad\qquad \varphi'' = \tanh^{-1}\frac{v''}{c}$$

these two successive Lorentz transforms

$$\begin{aligned}\Delta x' &= \cosh\varphi'(\Delta x - \tanh\varphi'\ c\Delta t)\\ c\Delta t' &= \cosh\varphi'(-\tanh\varphi'\ \Delta x + c\Delta t)\end{aligned}$$

and

$$\begin{aligned}\Delta x'' &= \cosh\varphi''(\Delta x' - \tanh\varphi''\ c\Delta t')\\ c\Delta t'' &= \cosh\varphi''(-\tanh\varphi''\ \Delta x' + c\Delta t')\end{aligned}$$

are overall equivalent to the transform

$$\begin{aligned}\Delta x'' &= \cosh\Phi(\Delta x - \tanh\Phi\ c\Delta t)\\ c\Delta t'' &= \cosh\Phi(-\tanh\Phi\ \Delta x + c\Delta t)\end{aligned}$$

where

$$\Phi = \varphi' + \varphi''$$

31. A point mass, being of infinite density, would actually constitute a black hole.

(a) At what rate, compared to a clock infinitely far from a point mass m, does time pass at twice the Schwarzschild radius $r = 2mG/c^2$?

(b) At what rate, compared to a clock infinitely far from a point mass m, does time pass at the Schwarzschild radius $r = 2mG/c^2$?

(c) What strange thing happens to the elements of the metric (10.99)

$$g_{\mu\nu} = \begin{pmatrix} 1-\dfrac{2mG}{rc^2} & 0 & 0 & 0\\ 0 & -\left(1-\dfrac{2mG}{rc^2}\right)^{-1} & 0 & 0\\ 0 & 0 & -r^2 & 0\\ 0 & 0 & 0 & -r^2\sin^2\theta \end{pmatrix}$$

as you pass from $r > 2mG/c^2$ to $r < 2mG/c^2$? And what's up with that?? (Note that a familiarity with matrices is not required to answer this question.)

10.12 Sketchy Answers

(1)

Part	Earth years	Enterprise years	Enterprise light-years
a	3600	3600	36.0
b	360	358	35.8
c	51.4	36.7	25.7
d	36.4	5.13	5.08
e	36.0	0.509	0.509
f	36.0	0.0509	0.0509

(2a) $\frac{4}{5}$.

(3a) $\frac{12}{13}c$. (You have a good arm.)

(3b) $\frac{3}{25}$ light-seconds.

(4(a)i) 4.0 sec and -4.0×10^8 m.

(4b) 6.0 sec and 14×10^8 m.

(6a) 168.1 yr and 164 ly.

(6b) 36.9 yr, and that's all you're getting.

(10a) $\dfrac{1}{\sqrt{1 + \left(\dfrac{cT}{\ell}\right)^2}}$.

(10b) $T\sqrt{1 + \left(\dfrac{\ell}{cT}\right)^2}$.

(11a) 91 yr.

(11c) 119 yr.

(11d) 96 yr.

(11e) 144 yr.

(12) 2 yr.

(15) 1.25×10^{-5}.

(16a) $1 + 1.6 \times 10^{-16}$ hr.

(16b) 12 mi. Don't you feel silly now?

(16c) $12 + 1.9 \times 10^{-15}$ mi.

(17a) $3.82063609 \times 10^{-12}$ J $= 23.8465364$ MeV.

(17b) 5.7118293×10^{14} J.

(20) $\frac{2}{5}c$.

(21) $\frac{7}{8}c$.

(22a) $\frac{2}{5}c$.

(22b) $\frac{10}{11}c$.

(22c) Dude!

(25) $\frac{3}{5}c$.

(26a) With $\beta_t = \dfrac{v_t}{c}$ and $\beta_b = \dfrac{v_b}{c}$, $\quad \dfrac{1-\beta_t^2\beta_b^2}{1-\beta_t^2}\,\dfrac{\ell}{2v_b}$.

(26b) With the same notation, the time between is $\dfrac{\beta_t}{\sqrt{1-\beta_t^2}}\,\dfrac{\ell}{c}$.

(26c) With the same notation, $\quad \dfrac{(1+\beta_t\beta_b)(1-2\beta_b\beta_t+\beta_t^2)}{(1-\beta_t^2)^{\frac{3}{2}}}\,\dfrac{\ell}{2v_b}$.

(28a) $(p_x, p_t) = (\gamma_0 m v_0, \gamma_0 mc)$ and $(p_x, p_t) = (0, mc)$. Not rocket science, but you can't get very far in this problem without these.

Chapter 11

Fluid Dynamics

11.1 The Bernoulli Equation

A fluid (that is, a gas or a liquid) is amorphous. Since the shape and even the volume occupied by a fluid can change, we cannot analyze the dynamics of fluids in the same way that we analyze the dynamics of rigid bodies; though fluids of course obey the same physical laws as rigid bodies, these laws must be expressed in a different form for them. Because the bits of fluid that make up any finite volume or mass can move independently of each other, we need instead to set up relations for *points* within the fluid, by looking at infinitesimal volumes and then taking the limit as these volumes shrink to zero size.

For a rigid body moving under the influence of gravity near the Earth's surface, for example, conservation of energy would take the form

$$\tfrac{1}{2}mv_1^2 + mgy_1 = \tfrac{1}{2}mv_2^2 + mgy_2$$

For a fluid, we want to apply this relation not to a finite mass m occupying a finite volume V, but to an infinitesimal mass dm occupying an infinitesimal volume dV:

$$\tfrac{1}{2}\,dm\,v_1^2 + dm\,gy_1 = \tfrac{1}{2}\,dm\,v_2^2 + dm\,gy_2 \tag{11.1}$$

Taking the limit of this as $dm \to 0$ would of course be just stupid: we would end up with $0 = 0$ — perfectly correct, but not terribly useful. To derive something useful from eq. (11.1), we need to express it in terms of some quantity or quantities that remain finite as $dm \to 0$ and $dV \to 0$, and that quantity would be the mass density (mass per unit volume) ρ:[1]

$$\rho = \frac{dm}{dV} \tag{11.2}$$

[1] In general, this mass density is a function of location $\mathbf{r}$ and time t, so that $\rho = \rho(\mathbf{r}, t)$: if the fluid is compressible, the density of the fluid may differ from one point to another or change as time passes. But at any given point and time, the density is finite.

Dividing both sides of eq. (11.1) by dV, we obtain

$$\frac{1}{2}\frac{dm}{dV}\,v_1^2 + \frac{dm}{dV}\,gy_1 = \frac{1}{2}\frac{dm}{dV}\,v_2^2 + \frac{dm}{dV}\,gy_2$$
$$\tfrac{1}{2}\rho v_1^2 + \rho g y_1 = \tfrac{1}{2}\rho v_2^2 + \rho g y_2 \tag{11.3}$$

which, in the limit $dV \to 0$, holds at each point in the fluid, with ρ being the density of the fluid at that point.

That is, eq. (11.3) almost holds. In reasoning from a relation governing the motion of rigid bodies we have, however, overlooked the contribution of a property unique to fluids. Unless you've been living in a closet all your life, you know that the pressure p in a fluid is, by definition, the outward force per unit area that the fluid exerts on whatever surrounds it: [2]

$$p = \frac{F}{A} \tag{11.4}$$

The MKS unit of pressure is the Pascal (Pa): $1\,\text{Pa} = 1\,\text{N/m}^2$. Another common unit of pressure is the atmosphere (atm): $1\,\text{atm} = 1.01325 \times 10^5\,\text{Pa}$, or, in English units, $14.70\,\text{lb/in}^2$. And as we will see later, a pressure may also be expressed in terms of the height of a column of mercury that it could support, so that $1\,\text{atm} = 760\,\text{mm Hg}$ and $1\,\text{mm Hg} = 133.3\,\text{Pa}$.

Pressure is relevant to our relation (11.3) for conservation of energy because the corresponding force would do work during an expansion or contraction of the fluid. To see how much work, consider a fluid enclosed in a cylinder of cross-sectional area A fitted with a movable piston.[3] As the fluid expands and pushes the piston outward a distance $d\ell$, the work dW done by the force F associated with the pressure is

$$dW = F\,d\ell$$

Using $F = pA$ from eq. (11.4), we can re-express this as

$$dW = pA\,d\ell = p\,dV$$

[2] Pressure, like density, need not be constant, so more strictly we should write eq. (11.4) in terms of the infinitesimal force dF exerted over an infinitesimal area dA, as $p = dF/dA$. But eq. (11.4) is good enough for our purposes.

Anyway, not only is there now a collision in notation between momentum p and pressure p, but you must also be careful to distinguish between p and ρ. Alas, so few letters in the alphabet, and so many quantities to denote! And while the p and ρ thing might seem gratuitously malicious, ρ is in fact the conventional symbol for density in all sorts of other contexts. So you'll just have to deal with it.

[3] Arguably redundant, as there are few applications for immovable pistons. But you get the idea.

where we have noted that $A\,d\ell$ is the corresponding infinitesimal increase dV in the volume of the fluid. The work per unit volume is therefore

$$\frac{dW}{dV} = p \tag{11.5}$$

Although we have derived this result only for the special case of an expansion into an infinitesimal cylindrical volume dV, an expansion into any finite volume of any shape can be built up out of such infinitesimal cylindrical volumes, so our result is quite general.

Recall now that our tentative relation (11.3) for conservation of energy in a fluid was expressed per unit volume. Since any work done by a fluid will detract from its energy, we must take this into account in our eq. (11.3) by including the work per unit volume $dW/dV = p$:

$$p_1 + \tfrac{1}{2}\rho v_1^2 + \rho g y_1 = p_2 + \tfrac{1}{2}\rho v_2^2 + \rho g y_2 \tag{11.6}$$

This final result for conservation of energy in fluids is known as the Bernoulli equation. Note that for fluids the pressure is a sort of potential energy: to compress the fluid you must exert a force on it, and the corresponding work you do during the compression is stored in the fluid in the form of its pressure. This "potential energy" can later be released by allowing the fluid to expand.

Most of our practical applications will deal with plumbing — that is, with relatively incompressible liquids like water traveling through cylindrical pipes. If a liquid is incompressible — as most common liquids in fact are to at least a good approximation — , then in a given time interval the same volume of liquid must flow past each point along the pipe, even if the pipe becomes wider or narrower. If A is the cross-sectional area of the pipe, and if the liquid at that point in the pipe is moving at speed v, then the distance the liquid moves down the pipe in a time interval dt is $d\ell = v\,dt$, and the volume of liquid that flows past is therefore $A\,d\ell = Av\,dt$. Since this volume must be the same at all points in the pipe for a given time interval dt, we have

$$\begin{aligned} A_1 v_1\,dt &= A_2 v_2\,dt \\ A_1 v_1 &= A_2 v_2 \end{aligned} \tag{11.7}$$

for any two points 1 and 2 along the pipe. For reasons that will remain obscure because we don't want to get into the vector calculus needed to express it in a more general form, eq. (11.7) is known as the equation of continuity. It is frequently useful in conjunction with the Bernoulli equation.

The quantity Av has its own simple physical interpretation: since $Av\,dt$ was the volume of fluid passing by each point in the pipe during the time interval dt, Av is therefore the flow rate (volume per unit time).

11.2 Archimedes's "Principle"

At this point, we should mention Archimedes's principle. It's not really much of a principle, but it does seem to have meant a lot to Archimedes. Archimedes's principle has to do with floating or submerged objects, and the gist of it is this: the buoyant force on an object is equal to the weight of the fluid displaced. To prove Archimedes's principle, simply consider a long rectangular block of cross-sectional area A that is submerged vertically to a depth h in a fluid of density ρ. The pressure in this fluid will exert a force over all of the submerged surface of the block. According to the Bernoulli equation (11.6), if the fluid is still (so that $v = 0$), then the pressure p at depth ℓ in the fluid is

$$p = p_0 + \rho g \ell$$

where p_0 is the pressure at the surface of the fluid, and where we have set $y = \ell$ at the surface and $y = 0$ at depth ℓ. Since the pressure therefore varies only with the depth ℓ, at any given depth the force exerted by the pressure on opposite sides of the submerged surface of the block is equal in magnitude but opposite in direction, so that there is no net contribution to the force on the block. The upward force exerted by the pressure on the bottom face of the block is

$$F_{bottom} = p_{bottom} A = (p_0 + \rho g \ell) A$$

If the top of the block is above the fluid, in a rarefied medium like the air, then the pressure on the top face of the block will have pretty much the same value p_0 that it does at the surface of the fluid. The downward force on the top face of the block will therefore be

$$F_{top} = p_{top} A = p_0 A$$

The net upward force on the block is thus

$$\begin{aligned} F_{buoyant} &= F_{bottom} - F_{top} \\ &= (p_0 + \rho g \ell) A - p_0 A \\ &= \rho g \ell A \end{aligned}$$

Now, ℓA, the product of the depth to which the block is submerged in the fluid with the cross-sectional area of the block, is just the submerged volume. So $\rho \ell A$ is just the mass of fluid corresponding to this volume (that is, the mass of fluid displaced by the block) and $\rho g \ell A$ is the weight of the displaced fluid.[4]

[4] What if we don't assume that the pressure on the top face of the block is the same as that at the surface of the fluid? In that case, you can separately calculate the buoyant forces on the submerged part of the block and on the part that projects above the surface,

If the top of the block is submerged in the fluid, then the pressure on the top face of the block will be

$$p_{top} = (p_0 + \rho g \ell')A$$

where ℓ' is the depth at which the top face is submerged. In this case the buoyant force will work out to

$$\begin{aligned} F_{buoyant} &= F_{bottom} - F_{top} \\ &= (p_0 + \rho g \ell)A - (p_0 + \rho g \ell')A \\ &= \rho g (\ell - \ell')A \end{aligned}$$

Since $(\ell - \ell')A$ is the submerged volume of the block, we again conclude that the buoyant force equals the weight of the displaced fluid.

So far, our proof is limited to the case of a rectangular block submerged vertically in the fluid. One can, however, construct any shape out of vertical rectangular blocks of infinitesimal cross-sectional area. All we need to do to extend the proof of Archimedes's principle to these aggregate bodies is to account for the internal portions of the sides of the infinitesimal blocks, where the blocks' surfaces are in contact, not with the fluid, but with other blocks. And that is easy: since the forces the pressure would have exerted on the abutting sides of any two such blocks would have been equal and opposite, there's no difference. Archimedes's principle therefore holds for bodies of arbitrary shape.

Archimedes's principle won't be of much concern to us, and the time we just spent on it is out of all proportion to its importance, but for what it's worth, there you have it.

11.3 Frisbees & Airplanes

Fig. (11.1) shows, in cross section, the air flow around a frisbee. The air that passes by the flat bottom of the Frisbee does so essentially freely: this air was at rest before the Frisbee came along, and pretty much remains so as the Frisbee passes by. But the air that passes over the top of the Frisbee has to "scoot by" the constriction presented by the Frisbee's dome-shaped top

with $\rho g \ell A$ being the buoyant force on the submerged part. The buoyant force on the part that projects above the surface will similarly be $\rho_{air} g \ell_{air} A$, where ρ_{air} is the density of air and ℓ_{air} the extent of block above the surface, so that the total buoyant force will be $\rho g \ell A + \rho_{air} g \ell_{air} A$.

But since ordinary fluids like water are about 1000 times as dense as air, neglecting the contribution from the buoyant force due to the air is in fact a very good approximation. Nor do you generally worry about such buoyant forces when you weigh objects on scales. Or if you do, you need to get a life.

Figure 11.1: Air Flow Around a Frisbee

and therefore acquires some velocity as the Frisbee passes by. If we think in terms of the Bernoulli equation and compare

$$p + \tfrac{1}{2}\rho v^2 + \rho g y$$

for the air going over and under the Frisbee, we see that although there is no significant difference in height y, the air going over the Frisbee has a higher velocity v and therefore, to compensate, a lower pressure p. Since $p = F/A$, this means that the air above the Frisbee exerts less force on the Frisbee than the air under it, with the result that there is a net upward force on the Frisbee. This upward force is called the lift and is what causes Frisbees to "float."

The propellers or jet engines of an airplane provide only forward thrust; the lift to keep the plane in the air is provided by the wings, which, like a Frisbee, are more or less dome-shaped on top and flat on the bottom. But in the case of an airplane, only part of the lift is provided by this effect; the rest — in fact, most — of the lift is generated by the wings being shaped and tilted so that air is also redirected downward by the wing, and by Newton's III law as the wing exerts this downward thrust on the air, the air exerts an equal and opposite upward thrust on the wings.[5]

11.4 Brazilian Soccer

Fig. (11.2) shows a cross section, as seen from above, of a soccer ball kicked with a spin. As when it passes over the dome-shaped top of a Frisbee, the air that passes around the soccer ball has to scoot by and thus acquires a substantial velocity. In the absence of any spin, the air would pass symmetrically around all sides of the ball, right, left, top, and bottom. But with the spin shown in fig. (11.2), friction between the ball and the air pulls the air going around the left side of the ball to a higher velocity and likewise hinders and slows down the air going around the right side of the ball. By the same

[5] And it is because of this that airplanes are able to fly upside down: by adjusting the flaps and what is rather bizarrely known as the *angle of attack* — the tilt of the plane due to the relative heights of the nose and the tail — , it is possible to generate a sufficient downward thrust of the air to support the plane in inverted flight.

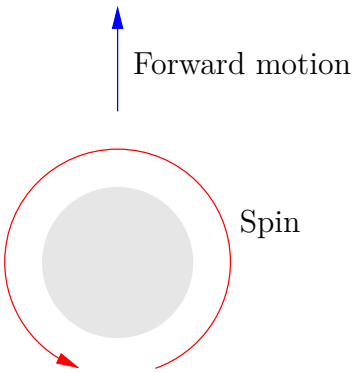


Figure 11.2: A Spinning Soccer Ball, Seen from Above

sort of reasoning as for the Frisbee, a higher air velocity around the left side means a lower pressure on the left side. The result is a net force to the left — which is in fact, as you know from common experience, the side to which the ball curves in flight.

Fig. (11.2) shows the clockwise spin that causes a ball kicked with the right foot to "hook" to the left. Similarly, a counterclockwise spin will cause the ball to "slice" to the right, topspin will give the ball a negative lift that causes it to drop very quickly, and backspin will give the ball a lift that causes it to rise (or at least not drop as quickly).

This same effect of course applies to the flight of any kind of ball: golf balls, curve-balls in baseball, etc. In fact in golf the lift due to backspin so significantly increases the range of the ball that the tilt of the heads of golf clubs and the patterns on their surfaces are designed to maximize it.

And while we're on the subject of golf ...

11.5 Why Golf Balls Have Dimples

In our work with fluids, we will always neglect turbulence — the swirling, whirlpool motion that tends to develop spontaneously as fluids flow around objects or, equivalently, as objects move through fluids. In addition to being very difficult to calculate mathematically, turbulence is often chaotic, in both the everyday and strict mathematical senses. Usually neglecting turbulence is a reasonable approximation, but there are situations in which its effect is very substantial. Such is the case with the flight of a golf ball: it turns out that a smoothly spherical golf ball would experience chaotic turbulence that would cause it to veer off in a random direction. While this would be very

entertaining for the spectators, it wouldn't make for very good golf.[6] The dimpling in the surface of the golf ball actually causes turbulence, but in a controlled, symmetric way, so that the ball does not veer off.

Turbulence is a frictional effect to which energy is lost: the energy of motion for the turbulent flow of the air comes from the energy of motion of the ball, which is therefore slowed down. But in the case of a golf ball, it turns out that the increase in friction around the sides of a dimpled ball is more than compensated for by a reduction in the turbulence that would occur in the wake behind a smooth ball, so that a dimpled ball actually travels faster and farther than a smooth ball.

Bullets experience similarly dramatic turbulence effects in flight, but the problem is resolved differently: rather than dimples, the bullet is given a spin by twisted grooves (called *rifling*) carved into the gun-barrel. This spin makes the bullet a miniature gyroscope and keeps its direction stable. So if ever they modify the rules of golf to allow shooting the ball out of a cannon, you can expect to see players using smooth balls and "golf guns" with rifled barrels.

In case you were curious, tennis balls have fuzz for the same reason that golf balls have dimples. And they would be even more fun to fire out of cannons.

[6] Of course, trying to propel a golf ball by swinging a stick at it while it's lying on the ground doesn't make for very good golf, either: if you were asked, as an engineering problem, to devise a way to get a golf ball from the tee to the hole, you would be very unlikely to come up with anything even remotely resembling a golf club. But that's another matter.

11.6 Problems

1. (a) In an article in a major New York newspaper that shall remain nameless, a clueless reporter once claimed that narrow high-heels, such as "stilettos," tend to punch holes in the carpet because they exert a greater force on the floor. Set this person straight.

   (b) How does the bed-of-nails trick work?

2. A major-league pitcher can throw a regulation 5 oz (142 g) baseball at you at nearly 100 mph (≈ 45 m/s). By comparison, the 158 grain (10.2 g) bullet from a .357 magnum leaves the barrel at 440 m/s. Compare the kinetic energy and momentum of the baseball to those of the bullet and explain the physical reason (as opposed to the physiological reason, with which we are all familiar) why getting shot is worse than getting hit with a fast-ball.

3. (a) How tall must a column of mercury (element Hg, density 13.6 g/cm^3) be in order for the pressure difference between the top and bottom to be 1 atm?

   (b) At what depth does the pressure in a column of water increase by 1 atm?

   (c) What is the pressure a mile under the sea, in atmospheres? (Sea water has an average density of about 1.03 g/cm^3.)

   (d) The word on the street is that your lungs can operate against a pressure difference of only about $\frac{1}{20}$ atm. What is the corresponding maximal depth at which you can snorkel?

   (e) How did they ever come up with a silly word like *snorkel*, anyway?

4. You have probably noticed that cracking the window when you are smoking in a moving vehicle will draw the smoke out through the window. Explain this effect in terms of the Bernoulli equation. (This same effect also helps you blow your nose.)

5. (Everything you ever wanted to know about Super Soakers, and then some.) Suppose the nozzle on your Super Soaker is 1.0 mm in diameter and that when you fire it horizontally over level ground from a height of 1.2 m the stream travels a horizontal distance of 6.6 m before hitting the ground.

   (a) Determine the speed at which the stream exits the nozzle of the Super Soaker.

   (b) Assuming that the reservoir is at approximately the same height as the nozzle and that the water in the reservoir will be moving so slowly that it is essentially at rest just before exiting the nozzle, determine the pressure in the reservoir of your Super Soaker.

   (c) In a typical Super Soaker, the reservoir is a few centimeters above the nozzle. Demonstrate quantitatively that neglecting the height difference between the reservoir and the nozzle is a good approximation. Is this still true if you have one of those models with tanks that strap onto your back?

   (d) Determine the rate (in kg/sec) at which the water is coming out of your Super Soaker.

   (e) Determine the force exerted on your friend if you blast said friend point blank in the face. For simplicity, assume that the water essentially comes to rest after hitting the target. See the footnote if you need a hint.[7]

   (f) Is there any recoil when firing a Super Soaker?

   (g) To spice things up, you fill your Super Soaker with gasoline, which is less dense than water, and use it as a flamethrower.[8] Will the Super Soaker shoot farther, not as far, or the same distance that it did when filled with water?

6. Remember those little water rockets — the ones you fill not quite full of water and then pump up? Explain how they work. What is the purpose of leaving the little air pocket? And why use water — why not just pump them full of air? If you need a hint, it might help to glance back at §6.6 on p.306.

7. A typical Frisbee is 175 g and 10.5 in (26.7 cm) in diameter. Work out a rough figure for the velocity at which the air "scoots over" the top of the Frisbee. The density of air is about $1.29\,\mathrm{kg/m^3}$.

8. Explain why Frisbees shift erratically up and down on days when there is a gusty wind.

[7] Think in terms of $F = dp/dt$.

[8] Bear in mind, of course, that were you actually to do this, you would likely end up being the one doing the entertaining (to the extent that the mingled odors of barbecued meat and burning plastic are entertaining).

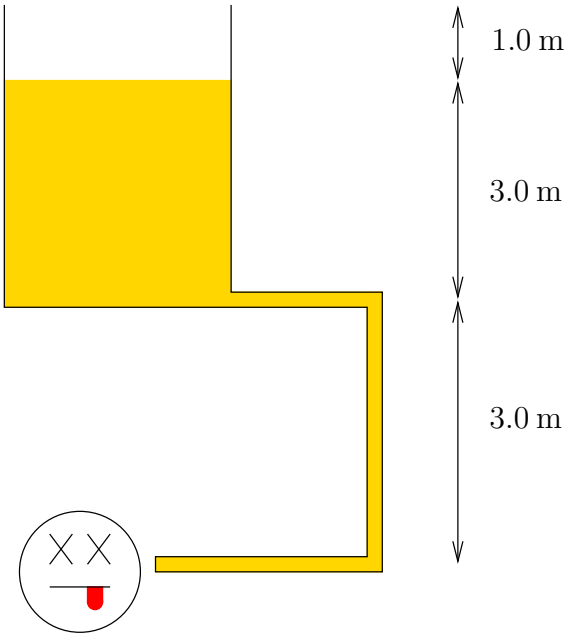


Figure 11.3: Problem 9

9. The purpose of a water tower is to generate pressure that will in turn increase the flow rate from faucets, etc.

   Suppose that, in order to ensure an adequate flow rate, on the roof of your dorm you rig up a similar tower to supply your favorite super-caffeinated, high-sugar, FDA Schedule I energy drink, with the dimensions shown in fig. (11.3).

   (a) At what speed will the drink (essentially the same density as water) exit the nozzle at the business end?

   (b) What diameter of nozzle would be required for a flow rate of 3.0 liters per second?

   (c) If the nozzle were replaced with a shower head, the drink would in fact come out of the shower head at a considerably lower speed than it would out of the much wider nozzle of #9a. What physical effect(s) have we neglected that would reduce the fluid's speed, and why is this effect more significant for a narrower aperture?

10. While smoking a cigar, you pour gasoline into an initially empty vessel that has, unknown to you, a hole of area A in its bottom.

    (a) If you continue pouring gasoline in at a constant rate, the depth of the gasoline will rise, but at a decreasing rate, and if the vessel is large enough the depth will eventually level off at some maximal value. Why does this happen?

    (b) Determine this maximal depth if you are pouring the gasoline in at a constant rate Q, where Q is volume per unit time.

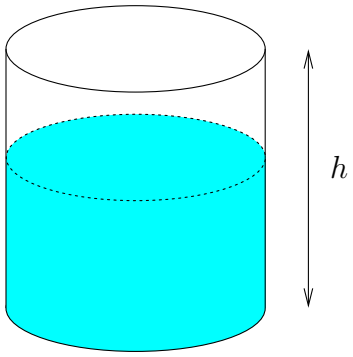


Figure 11.4: Problem 11

11. Just to see what will happen, you punch a hole in the bottom of a drum of radioactive sulfuric acid. The drum, of radius R, is open at the top and initially filled to its full height h with the acid (density ρ). The hole (not shown in fig. (11.4)) is of radius r. Determine the total time it will take the drum to empty.

12. You tie your little brother/sister to a chair and tape up his/her face so that he/she has to breathe through a straw. The straw has a diameter of 5.0 mm. The density of air is about $1.29\,\mathrm{kg/m^3}$.

    (a) What is the greatest speed at which air can be "sucked" through the straw? Remember that human lungs can generate a pressure difference of only about $\frac{1}{20}$ atm.

    (b) Why is it not strictly correct to speak of the air being *sucked* through the straw? What is actually happening physically?

    (c) At what rate (in liters per second) is it therefore possible to breathe through the straw?

    (d) Will the sibling survive? You should be able make a reliable enough quantitative estimate to draw a conclusion — just out of your head, without looking anything up.

    (e) In fact, the speed at which air could be drawn through the straw would be quite a bit less than what you just calculated. What physical effect(s) have we neglected that would reduce the speed?

13. Inspired by the Black Knight in *Monty Python & the Holy Grail*, you take a big, sharp sword, sneak up behind your roommate, and lop his or her head cleanly off.[9]

    Suppose that your roommate (at least up to this point) had a fairly normal blood pressure of 120/80 in the conventional units of mm Hg, that is, millimeters of mercury.[10] The systolic reading, 120, is the peak pressure with each heartbeat; the diastolic reading, 80, is the background blood pressure between beats; and both readings specify the pressure difference, that is, the pressure above the ubiquitous atmospheric pressure of 760 mm Hg.

    (a) With each heart beat, how high does blood spurt straight vertically upward from the neck? Assume that your roommate remains momentarily upright and that the blood within the neck is essentially at rest before being spurted out. The density of blood varies with, among other things, one's state of hydration, but at body temperature is about $1.05\,\mathrm{g/cm^3}$.

    (b) You have probably never beheaded a roommate, but you have certainly cut yourself at some point and have noticed that you do not, as your answer to the preceding part would seem to suggest, bleed like a geyser. What physical or physiological effects spoil the fun?

    (c) Your blood pressure is usually taken in your upper left arm, closest to and on a level with your heart. If your blood pressure is 120/80 at the level of your heart, what is it in your feet, which we will take to be 1.25 m below your heart?[11]

14. You may have noticed that a "nice" stream of tap water — one that isn't chaotically turbulent — narrows as it descends. Then again, you may be a totally unobservant person who's never noticed this. But now you know. Anyway, show that the diameter of such a vertical stream that exits the faucet at speed v_0 narrows by a factor of

$$\left(1+\frac{2gh}{v_0^2}\right)^{\frac{1}{4}}$$

    as it descends a distance h.

[9] *'Tis only a flesh wound!*

[10] Actually, normal blood pressure should be about 100/60; 120/80 is only considered normal in industrialized countries where people lead very unhealthy lifestyles.

[11] This is, of course, neglecting some of the effects cited in #13b.

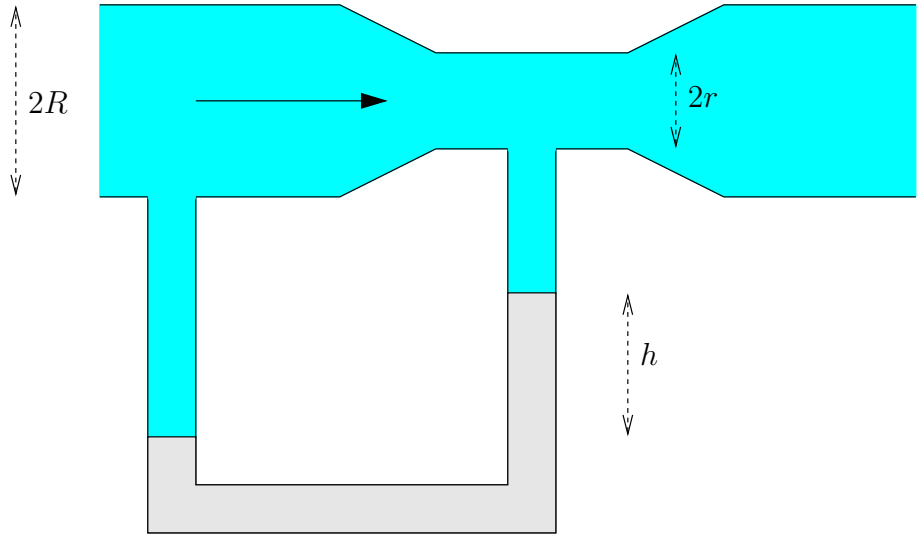


Figure 11.5: Problem 15

15. Fig. (11.5) shows a fluid (light blue, of density ρ) flowing through a cylindrical tube of variable radius in the direction of the solid arrow. The wide part of the cylindrical tube has radius R, the narrow part radius r. Underneath this cylindrical tube, and opening into it, is another tube containing a different fluid (gray, of density ρ', and of course denser than and immiscible with the blue fluid). All the fluid in this underlying U-shaped tube is effectively stationary, and the difference in height of the gray fluid between its two sides is h.

    Now that we've gotten all that out of the way, determine the speed at which the water is moving through the wide part of the cylindrical tube.

    (This whole arrangement is known as a Venturi tube and can, as you have just found, be used to determine the velocity of a fluid. The U-shaped tube underneath by itself is known as a manometer and is used to measure pressure differences.)

16. Oils and fats are typically only about 70% as dense as water. Explain how weighing a person in air and then again in a pool filled with water can be used to derive a result for that person's percentage of body fat. (One could, of course, go for a more direct measurement of a person's density by weighing the person in air and then dunking him or her in a large graduated cylinder, but, while arguably more entertaining, this unfortunately turns out not to be as practical.)

11.7 Sketchy Answers

(3a) 0.760 m.

(3b) 10.3 m.

(3c) 157.

(3d) 0.50 m in salt water, 0.52 in fresh.

(5a) 13 m/s.

(5b) 1.88 atm.

(5d) 0.0105 kg/sec.

(5e) 0.14 N.

(9a) 11 m/s.

(9b) 1.9 cm.

(10b) $\dfrac{Q^2}{2gA^2}$.

(11) $\dfrac{R^2}{r^2}\sqrt{\dfrac{2h}{g}}$.

(12a) 89 m/s.

(12c) 1.7 liter/sec.

(13a) 1.5 m.

(13c) About 220/180.

(15) $\sqrt{\dfrac{\rho' - \rho}{\rho}\, 2gh\, \dfrac{r^4}{R^4 - r^4}}$.

Chapter 12

Things ~~Nukuler~~ Nuclear

The word is Nū′-klē-ər, not Nū′-kyə-lər. Do you really want to go around sounding like George Bush?

12.1 The Composition of Nuclei

Unless you grew up in an extremely sheltered environment, you know that atoms consist of negatively charged electrons orbiting around a positively charged nucleus. The nucleus in turn is composed of two kinds of particles generically known as nucleons: protons, which are positively charged, and neutrons, which are uncharged. In terms of the fundamental charge $e = 1.60217653 \times 10^{-19}\,\mathrm{C}$, the electron carries charge $-e$ and the proton charge $+e$. Although for most purposes it is sufficient to regard the nucleus as a collection of nucleons, it turns out that on a finer scale protons and neutrons are in turn composed of still smaller particles known as quarks, predominantly the so-called up and down quarks.[1] These up and down quarks are denoted symbolically by u and d and carry charges $+\frac{2}{3}e$ and $-\frac{1}{3}e$, respectively. Each proton turns out to consist of two up and one down quark (uud) and each neutron of two down and one up quark (udd). In terms of quark charges, the charge on the proton and neutron thus resolve into

$$\begin{aligned} 2 \times \tfrac{2}{3}e - \tfrac{1}{3}e = +e \quad &\checkmark \\ 2 \times -\tfrac{1}{3}e + \tfrac{2}{3}e = 0 \quad &\checkmark \end{aligned}$$

Other properties of the proton and neutron are similar composites of the properties of their constituent quarks. These constituent quarks are held

[1] This terminology has to do with a property known as nuclear isospin, which is very similar to the up and down spin angular momentum of electrons with which you are no doubt familiar from a previous life in chemistry.

together by the strong force, so called because it is stronger than the electromagnetic force that would otherwise blow the nucleus apart due to the electrical repulsion of its positively charged protons.

When the nucleus is regarded as a collection of protons and neutrons, it is categorized by three parameters:

$$\begin{aligned} \text{atomic number } Z &= \text{number of protons} \\ N &= \text{number of neutrons} \\ \text{atomic mass } A &= \text{total number of nucleons} \\ &= N + Z \end{aligned}$$

In an atom the number of electrons orbiting around the nucleus will equal the number of protons in order to balance the overall charge. The atomic number Z of a nucleus therefore determines to which chemical element the nucleus corresponds. For a given Z there are, however, nuclides of several differing values of the number N of neutrons. Nuclides of the same Z but differing N are known as isotopes: the differing values of N of course make them different nuclides, but they all correspond to the same element, and their atoms are virtually identical chemically. All carbon nuclei, for example, have six protons, but the number of neutrons varies from four to nine, with carbon-12 (six protons and six neutrons) being by far the most common isotope.

There are several different notations in use: for a nucleus of element X, you may see any of

$$^{A}\mathrm{X} \qquad\qquad ^{A}\mathrm{X}_{Z} \qquad\qquad {}^{A}_{Z}\mathrm{X}$$

We will generally use ${}^{A}_{Z}\mathrm{X}$. For example, the most common isotope of helium is helium-4, which consists of two protons and two neutrons. Thus $Z = 2$, $N = 2$, $A = 2 + 2 = 4$, and we would write ${}^{4}_{2}\mathrm{He}$.

Protons and neutrons differ slightly in mass, and the potential energy associated with the force binding the nucleons together also alters the mass of a nucleus according to $E = mc^2$: the higher the binding energy, the lower the potential energy and thus mass of the nucleus. One conventional measure of mass, the atomic mass unit (u), is defined to be one twelfth of the mass of a carbon-12 nucleus:

$$1\,\mathrm{u} = 1.66053886 \times 10^{-27}\,\mathrm{kg}$$

The masses of isolated protons and neutrons are[2]

$$\begin{aligned} 1.67262171 \times 10^{-27}\,\mathrm{kg} &= 1.00727647\,\mathrm{u} \qquad \text{proton} \\ 1.67492728 \times 10^{-27}\,\mathrm{kg} &= 1.00866492\,\mathrm{u} \qquad \text{neutron} \end{aligned}$$

[2]Here you can see the effect of the binding energy on nuclear mass: for carbon-12, the naive sum of six proton and six neutron masses would give $6(1.00727647)+6(1.00866492) = 12.0956483\,\mathrm{u}$ instead of $12\,\mathrm{u}$.

12.2 Types of Nuclear Decay

When nuclear decay and radiation were first observed, the mechanism of the decay and the kinds of matter that constituted the various forms of radiation were a mystery, and for lack of any more informed scheme people simply enumerated the various kinds of observed decays according to the letters of the Greek alphabet: α, β, and γ. Had a fourth decay scheme been observed, it would have been called δ decay.[3]

Alpha decay In α decay, a large nucleus spits out a helium-4 nucleus — the α particle — , which consists of two protons and two neutrons. Such decays occur spontaneously because the potential energy of the system decreases: the total potential energy of the α particle and the product nucleus is lower than that of the parent (that is, the original) nucleus. Since the α particle carries away two protons and two neutrons, the number of protons in the parent nucleus is reduced by two and the atomic mass by four:[4]

$$ {}^{A}_{Z}\mathrm{X} \longrightarrow {}^{A-4}_{Z-2}\mathrm{Y} + {}^{4}_{2}\alpha $$

Note that since the product nucleus has a different atomic number, it corresponds to a different element, so we have used the generic symbol Y for it to distinguish it from the parent nucleus X.

Beta decay In β decay, the nucleus spits out a β particle (an electron, symbol e^-)[5] and an uncharged particle known as an antielectron neutrino (symbol $\bar{\nu}_e$).[6] When this happens, a neutron inside the nucleus turns into a proton: as the negatively charged electron is ejected, the nucleus acquires an equal quantity of additional positive charge by virtue of the new proton, thus conserving electric charge overall. On the level quarks, one of the down quarks in a neutron changes into an up quark by emitting a particle known as a W^- (read "W minus"), which, before

[3] Actually, it was: early on the terms δ ray and ϵ ray were used by some people for heavy recoil nuclei and tertiary radiation sources. But these terms didn't stick.

[4] Sometimes the ${}^{4}_{2}$ is omitted from the α, or, since it is in fact a helium nucleus, the ${}^{4}_{2}\alpha$ is written ${}^{4}_{2}\mathrm{He}$:

$$ {}^{A}_{Z}\mathrm{X} \longrightarrow {}^{A-4}_{Z-2}\mathrm{Y} + \alpha \qquad\qquad {}^{A}_{Z}\mathrm{X} \longrightarrow {}^{A-4}_{Z-2}\mathrm{Y} + {}^{4}_{2}\mathrm{He} $$

[5] The negative sign is to indicate the negative charge on the electron and distinguish it from its positively charged antiparticle, the positron (e^+).

[6] It is conventional to denote the antiparticle of an uncharged particle by putting a bar over the symbol for it: thus the ν for a neutrino becomes $\bar{\nu}$ for an antineutrino. Each of the three varieties of lepton — the electron e, the muon μ, and the tau τ — has its own associated kind of neutrino (ν_e, ν_μ, and ν_τ). Just in case you were curious.

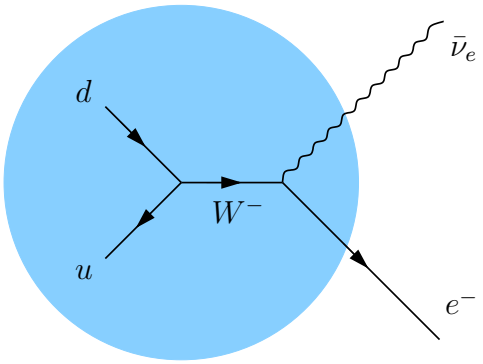


Figure 12.1: Beta Decay

traveling far enough to escape from the nucleus, decays into the electron and antielectron neutrino that are subsequently seen, as shown in fig. (12.1). The overall result in terms of quarks is

$$d \rightarrow u + e^- + \bar{\nu}_e$$

In terms of nucleons,

$$udd \rightarrow uud + e^- + \bar{\nu}_e$$
$$\text{neutron} \rightarrow \text{proton} + e^- + \bar{\nu}_e$$

In the context of nuclear decay, the ejected electron is sometimes denoted by ${}^{0}_{-1}e$ for the purposes of balancing charge: since changing a neutron into a proton increases the number of protons in the nucleus by one without changing the total number of nucleons, the general β-decay scheme is

$${}^{A}_{Z}\mathrm{X} \rightarrow {}^{A}_{Z+1}\mathrm{Y} + {}^{0}_{-1}e + \bar{\nu}_e$$

Note that, as in α decay, the product nucleus differs in atomic number from the parent nucleus and thus corresponds to a different element.

There are actually two kinds of β decay. The above scheme, in which an electron is ejected from the nucleus, is called β^- (read "beta minus") decay. The other type, β^+ ("beta plus") decay, comes in two flavors: A proton may change in to a neutron, with the emission of a positron (e^+, the positively charged antiparticle of the electron) and an electron neutrino (ν_e):

$${}^{A}_{Z}\mathrm{X} \longrightarrow {}^{A}_{Z-1}\mathrm{Y} + e^+ + \nu_e$$

This positron then quickly annihilates with an electron, giving rise to a γ ray. Alternatively, the nucleus may undergo electron capture and swallow up one of the atomic electrons orbiting around it. The overall

result is that this electron combines with a proton in the nucleus to produce a neutron and an electron neutrino:

$$e^- + {}^{A}_{Z}\mathrm{X} \longrightarrow {}^{A}_{Z-1}\mathrm{Y} + \nu_e$$

β^+ decay is rare, however, and we won't be dealing with it any further.

Neutrinos and their antiparticles, antineutrinos, are uncharged, massless particles that travel at the speed of light. Gazillions of them permeate the universe, but they interact so seldom with other forms of matter that you don't notice them. We'll say a little bit more about neutrinos and how they fit into the grand scheme of things in Chapter 22.

Gamma decay A γ particle (a.k.a. γ ray) is just a high-energy photon, that is, a high-energy light particle. An atom has energy levels corresponding to the orbital states of the electrons orbiting the nucleus; an electron can drop down from a higher to a lower orbit by emitting a photon, leaving the atom in a lower energy state. Although its nucleons orbit around each other rather than around a common center of attraction, a nucleus has very similar energy levels, and when it has been left in an excited state by an α or β decay or other nuclear process, it can drop down to a lower energy state by emitting a photon. The difference between the photons produced by atomic transitions and those produced by nuclear transitions is only the energy: the photons produced by nuclear transitions are typically a million times more energetic than those produced by atomic transitions.

Since the emission of a photon changes neither the number of protons nor the number of neutrons, the γ-decay scheme is very simple:

$${}^{A}_{Z}\mathrm{X} \longrightarrow {}^{A}_{Z}\mathrm{X} + \gamma$$

Remind me and we'll set up a miniature cloud chamber with some alcohol vapor and dry ice, so that you can actually see the trails left by decay particles as they shoot out of nuclei.

12.3 Decay Rates & Constants

Many isotopes are stable. The decay of those that are radioactive is governed by quantum theory, according to which it is impossible to know exactly when a given nucleus will decay and we can calculate only the probability that it will decay during a given time interval: each radionuclide has its own half-life, the interval of time during which a nucleus has a 50-50 chance of decaying. For a single nucleus, there is therefore a great deal of uncertainty

about when the decay will occur. And as long as the number of nuclei is small, there will still be significant statistical fluctuations in the number that decay during each half-life. But as the sample size grows, these fluctuations become proportionally smaller and less significant. It can be shown that for large samples, the proportional fluctuations in the number N of decays go as $1/\sqrt{N}$: if, during some interval of time, 10^6 decays are expected on average, the statistical deviations from this average will be on the order of $1/\sqrt{10^6} = 10^{-3} = 0.1\%$. In the limit of large N, we can therefore neglect the statistical fluctuations and treat the expected averages for the number of decays as exact.

If we denote half-life by τ, the number of half-lives that elapse during a time interval t is t/τ. Since the initial number N_0 of radioactive nuclei is reduced by a factor of $\frac{1}{2}$ during each interval of a half-life, the number N remaining after time t will be

$$N = N_0\left(\tfrac{1}{2}\right)^{t/\tau} = N_0\, 2^{-t/\tau} \tag{12.1}$$

Note that as long as we are consistent, the quantities N and N_0 of the isotope can be measured in any units — number of nuclei, grams, imperial quarts, whatever.

We can also think in terms of the rate of decay: this rate, dN/dt, should be proportional to the quantity of radioactive material remaining at any given instant. If we denote the constant of proportionality by λ, we therefore have

$$\frac{dN}{dt} = -\lambda N \tag{12.2}$$

where the negative sign takes into account that the quantity N of radioactive material is decreasing, so that dN and hence dN/dt are negative. The constant of proportionality λ is known as the decay constant, and the differential equation (12.2) is easily solved by separating variables and integrating: if the number of nuclei goes from N_0 at time $t = 0$ to N at time t,

$$\begin{aligned}
\frac{dN}{N} &= -\lambda\, dt \\
\int_{N_0}^{N} \frac{dN}{N} &= -\int_0^t \lambda\, dt \\
\ln N \Big|_{N_0}^{N} &= -\lambda t \Big|_0^t \\
\ln \frac{N}{N_0} &= -\lambda t
\end{aligned}$$

and therefore

$$N = N_0\, e^{-\lambda t} \tag{12.3}$$

Eqq. (12.1) and (12.3) are completely equivalent. In fact, by equating these two results for N, we arrive at a relation between the half-life τ and the decay constant λ:

$$\begin{aligned} N_0\, 2^{-t/\tau} &= N_0\, e^{-\lambda t} \\ \ln 2^{-t/\tau} &= \ln e^{-\lambda t} \\ \frac{-t}{\tau}\, \ln 2 &= -\lambda t \\ \lambda &= \frac{\ln 2}{\tau} \end{aligned} \tag{12.4}$$

The rate of decay λN is called the activity (short, of course, for *radioactivity*) and is usually specified in either Becquerels (Bq) or Curies (Ci):[7]

$$1\ \text{Bq} = 1\ \text{decay/sec} \qquad\qquad 1\ \text{Ci} = 3.7 \times 10^{10}\ \text{decays/sec}$$

Specifying the rate of decay in Becquerels or Curies is the proper way to report the intensity of a radiation leak.

When doing decay calculations you will need to remember a little bit of chemistry involving Avogadro's number. For example, suppose you want to calculate the activity of 1.0 g of pure carbon-14 (half-life 5700 yr) in Curies. By eq. (12.4),

$$\lambda = \frac{\ln 2}{\tau} = \frac{\ln 2}{5700} = 1.22 \times 10^{-4}\ \text{yr}^{-1}$$

The decay rate is thus

$$\lambda N = (1.22 \times 10^{-4})(1.0) = 1.22 \times 10^{-4}\ \text{g/yr}$$

To get the activity in Curies, we want to convert this to decays/sec. The conversion from years to seconds you should be able to handle. To relate grams to number of nuclei, you need to remember that there are Avogadro's number of nuclei for every 14 g of carbon-14:

$$1.22 \times 10^{-4}\ \frac{\text{g}}{\text{yr}} \times \frac{6.02 \times 10^{23}\ \text{nuclei}}{14\ \text{g}} \times \frac{\text{yr}}{3.16 \times 10^{7}\ \text{sec}} = 1.66 \times 10^{11}\ \frac{\text{decays}}{\text{sec}}$$

$$1.66 \times 10^{11}\ \frac{\text{decays}}{\text{sec}} \times \frac{1\ \text{Ci}}{3.7 \times 10^{10}\ \text{decays/sec}} = 4.5\ \text{Ci}$$

[7] In problem #4 you will be initiated into the mysteries of this seemingly odd definition of a Curie.

12.4 Fission & Fusion

In fission, a large nucleus is split into smaller nuclei; in fusion, smaller nuclei are combined to form a larger nucleus. In both cases, the total mass of the products is less than the total mass of the parent nuclei and the lost mass is converted into energy according to the relativistic relation $E = mc^2$.

Fissile nuclides, which split spontaneously only at low rates, can be induced to split by bombarding them with neutrons. The material used in early fission bombs, then referred to as atomic bombs or A-bombs even though the effect was of course nuclear rather than atomic, was uranium-235, which becomes unstable after absorbing a neutron and splits into two large "pieces" and several loose neutrons. These loose neutrons may be absorbed by neighboring uranium-235 nuclei, which then also split, thereby causing a chain reaction and, by virtue of the sudden release of large amounts of energy, an explosion. An aggregation of uranium large and dense enough to undergo a runaway chain reaction is called a critical mass. Critical mass is achieved in bombs by two basic methods: either a gun-like mechanism fires two sub-critical masses together, or a jacket of chemical explosives is detonated around a sub-critical mass of fissile material to compact it to a critical density (the so-called implosion mechanism). In modern fission bombs the uranium-235 has been superseded by plutonium-239 artificially produced in a kind of nuclear reactor known as a breeder reactor. Plutonium is very fissionable and readily produced by bombarding the relatively abundant uranium-238 isotope with neutrons, but is also much more difficult to work with because of its complex crystal structures and chemical behavior. Although the half-life of plutonium-239 is about 24,000 yr, uncertainties about how its structure ages on much shorter time scales are presently the cause of a great deal of hair-pulling by the United States military as it tries to assess the state of its arsenal without resorting to test detonations.

Nuclear reactors also work by fission. The difference between a bomb and a reactor is that in the reactor there are mechanisms for keeping the fission process under control by, for example, inserting control rods of cadmium or other materials to absorb loose neutrons and thereby limit the chain reaction. Although the fuel used by fission reactors is not enriched enough to result in a nuclear explosion, failure to keep the reaction under control can result in a partial or total meltdown: the nuclear fuel in the core of the reactor could become so hot that it would melt and possibly burn through the containment structure. Fires or steam explosions could then disperse large quantities of radioactive material into the atmosphere and result, depending on wind and weather, in very widespread contamination. A slight partial meltdown occurred at the Three-Mile Island power-plant in Pennsylvania in 1979. In 1986, a much more serious meltdown occurred at the Chernobyl power-plant in the Soviet Union. Although very little radioactive material escaped from

Three-Mile Island, the Chernobyl disaster resulted in a massive release that spread over many countries.

Uranium occurs naturally; it is one of the elements that have made up the Earth since it first coalesced out of space dust, and it can be mined from veins of ore just as gold and other metals are. This uranium is still around because it has an extremely long half-life: 700 million years for uranium-235 and 4.5 billion years for uranium-238. Of these two isotopes, uranium-238 is vastly more common, but only uranium-235 is sufficiently fissile for use in bombs and reactors. Because the two isotopes are chemically identical and differ by less than 1% in mass, separating them from each other to obtain enriched uranium-235 is technically very difficult. The two most common methods are gaseous diffusion and centrifuging. In gaseous diffusion, uranium hexafluoride gas is allowed to migrate through a series of porous barriers. Since the molecules with uranium-235 are slightly lighter than those with uranium-238, on average they move slightly faster and therefore migrate slightly more quickly through the series of barriers. Alternatively, the uranium hexafluoride gas can be spun in a centrifuge: the slightly heavier uranium-238 molecules settle toward the outside and the enriched uranium toward the inside is skimmed off. By either process, the degree of enrichment is very slight, and the diffusion or centrifuging must be repeated many times to obtain sufficiently enriched uranium-235. It is principally the difficulty of separating uranium that has to date limited the number of nuclear powers in the world.

The primary difficulty in fusion is overcoming the electrical repulsion between nuclei: it takes a great deal of energy to get them close enough to each other for the strong nuclear force, which is of very short range, to become dominant and fuse them together. Fusion occurs naturally at the extremely high temperatures at the cores of active stars, and in fact all elements higher than hydrogen were originally produced by stellar fusion. To minimize the electrical repulsion and thus the energy required to cause fusion, terrestrial reactors use only the lighter elements as fuel, principally isotopes of hydrogen, helium, and lithium. Ordinary hydrogen nuclei consist of a single proton (${}^{1}_{1}\mathrm{H}$), but there are two isotopes:[8] deuterium, which consists of one neutron and one proton (${}^{2}_{1}\mathrm{H}$), and tritium, which consists of two neutrons and one proton (${}^{3}_{1}\mathrm{H}$). Tritium undergoes β decay with a half-life of about 12 years. Since the tiny amount of tritium naturally produced by cosmic rays decays away too quickly to accumulate, tritium has to be produced artificially, by bombarding deuterium or lithium with neutrons. Deuterium is stable, and although it constitutes only a tiny proportion of the hydrogen around us, the Earth's oceans, with their vast quantities of water, provide an essentially limitless supply of it in the form of heavy water — $\mathrm{H_2O}$ in which

[8] Three, if you count the quadium of Leonard Wibberly's *The Mouse That Roared*.

one or both of the hydrogens is deuterium, and which is so called because its molecular weight is therefore heavier than that of water containing only ordinary hydrogen. A variety of fusion reactions involving ordinary hydrogen, deuterium, tritium, helium-3, and lithium-6 are possible, and the products of these reactions are principally hydrogen, deuterium, tritium, helium-3, and helium-4. For the purposes of power generation, the fuel for the most promising fusions are hydrogen-boron,

$$ {}^{1}_{1}\mathrm{H} + {}^{11}_{5}\mathrm{B} \rightarrow 3\ {}^{4}_{2}\mathrm{He} $$

deuterium-tritium,

$$ {}^{2}_{1}\mathrm{H} + {}^{3}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n} $$

and deuterium-deuterium, which can produce either tritium and hydrogen or helium-3 and a loose neutron:

$$ 2\ {}^{2}_{1}\mathrm{H} \rightarrow \begin{cases} {}^{3}_{1}\mathrm{H} + {}^{1}_{1}\mathrm{H} \\ {}^{4}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n} \end{cases} $$

Fusion reactors use electromagnetic fields and sometimes lasers to heat and confine the fuel so that fusion can occur. For power production, they would be much cleaner and safer than the fission reactors currently in use: while fission reactors must be carefully controlled to prevent a runaway chain reaction resulting in a meltdown, and even when properly run produce highly radioactive waste that is difficult to dispose of safely, the worst catastrophe a fusion reactor could experience would be to simply stop working, and its waste could be used to fill party balloons in the parking lot.[9] Unfortunately, although there are working fusion reactors, they are not yet economically viable because, in addition to other technical engineering problems that remain to be solved, the energy required to cause the fusion to occur is still too high in comparison with the energy released. It is quite possible that further research would lead to fusion reactors that are economically viable, but somehow it has been judged more important to fund science-fiction fantasies like defense systems designed to shoot down mythical missiles launched by enemies that no longer exist than to try to solve the very real problem of energy production in a way that would be of immense benefit both economically and environmentally and that would vastly improve the quality of all of our daily lives.[10]

[9] This is actually a slight exaggeration, but only slight: because of the engineering technology required to run a fusion reactor, there is a risk of a fire or an explosion, but these would be of a conventional, not nuclear, nature and magnitude. And although the loose neutrons produced by fusion reactions can easily be shielded by a variety of common materials, some of these loose neutrons would inevitably form radionuclides when absorbed by nuclei in the internal structures of the reactor. The level of radioactivity involved would, however, be low enough that it would arguably not be a major health or environmental concern.

[10] Not that we have a definite opinion on the matter.

In 1989, a couple of chemists (*ugh!*) reported a discrepancy in energy when heavy water was hydrolyzed with palladium electrodes in a calorimeter. They attributed this discrepancy to fusion of the deuterium, and since this was happening at room temperature, the effect was dubbed cold fusion. It was hypothesized that there was something special about the electrodes that brought the deuterium nuclei close enough together for a long enough time that there was a significant probability of their fusing. The actual mechanism of the fusion would have been what is known as quantum tunneling,[11] a weird but well-established quantum process that, in this case, would allow the deuterium nuclei to combine with each other even though at room temperature they lacked the energy to overcome their mutual electrical repulsion. Cold fusion would thus have obtained the energy released by the fusion essentially for free. There was great excitement at the time, because the mechanism not only seemed quite plausible, but even to have the potential to solve the world's energy problems overnight: the metals involved were not uncommon, and the oceans contain a virtually limitless supply of deuterium. There was the prospect of myriad cold fusion generators providing all of the power needed both by society and by individuals at almost no cost beyond that of initially producing the generators. It even seemed possible that each home and car might be powered by its own shoe-box–sized generator. There would have been limitless power available for industrial production, recycling, desalination of sea-water, and all of humanity's other needs. And this power would have been not only essentially free, but totally nonpolluting: the only waste product would have been helium. Problems like the green-house effect, acid rain, nuclear waste, etc., would have been almost instantly relegated to history books.

Unfortunately, the whole thing ended up being a lesson in how premature announcements can confound the integrity of the scientific process. No one else was able to reproduce the results, and it quickly became clear that no fusion was in fact occurring. Research continues — quantum tunneling is well established and cannot be ruled out as a possible mechanism for bringing about fusion at low temperatures — , but there has still been no success in producing cold fusion.

There has, however, been great success producing fusion in weapons. Also known as a hydrogen bomb or H-bomb, a thermonuclear weapon is actually a hybrid in which the high temperature needed for fusion is generated by fission: At the core of a thermonuclear weapon, within a jacket of lithium deuteride ($^{6}_{3}\mathrm{Li}\,^{2}_{1}\mathrm{H}$),[12] is a plain old-fashioned fission bomb. The energy and the pressure wave produced by the fission bomb heats and compresses the jacket enough

[11] We will cover quantum tunneling in §22.5.3.

[12] Lithium deuteride is used, rather than pure deuterium or tritium, because it is solid at normal temperatures and because the lithium turns out to contribute greatly to the fusion process.

to cause fusion. This jacket of fusible material may also be surrounded by a second, outer jacket of fissionable material such as uranium-238: even though uranium-238 cannot undergo a chain reaction on its own, the loose neutrons produced by the fusion reaction induce fission of the uranium. The overall result is that a thermonuclear weapon releases vastly more energy, and consequently a blast many times more powerful, than a pure fission weapon. And as a bonus, thermonuclear weapons are more economical because they can be constructed with a smaller quantity of highly fissionable material.

Nuclear weapons are lethal in several ways. First, the tremendous energy suddenly released at detonation creates a giant fireball and shock wave. While the shock wave causes physical destruction like the collapse of buildings in the square miles near the blast, the heat from the fireball is actually more devastating and of longer range. The fireball emits thermal electromagnetic radiation, just like the thermal radiation you feel from a bonfire or red-hot poker, but intense enough to set buildings and other objects on fire several miles away. Nuclear radiation emitted immediately upon detonation is called prompt radiation. Much of this radiation is absorbed by earth, objects, and air before it gets very far, but there is enough to cause radiation sickness or death in people unfortunate enough to have survived the shock wave and thermal radiation. The fission and fusion processes also produce a variety of radionuclides. Many of the direct products of the fission process — the "daughter" nuclei into which the uranium or plutonium nuclei split — are themselves radioactive. Those with long half-lives continue to be a source of dangerous levels of radioactivity long after the blast. In addition, some previously innocuous nuclei in matter near the blast are rendered radioactive by absorption of stray neutrons or collisions with them. Ashes and other debris contaminated with these isotopes, drawn high into the atmosphere by the fireball and then carried by the winds, may fall out as snowflake-like ash or contaminated rain-drops. The more minute particles may be very widely distributed. Of particular concern are cesium-137 (half-life 30 yr); strontium-90 (half-life 29 yr), which behaves chemically like calcium and thus accumulates in the bones; and iodine-131 (half-life 8 days), which accumulates in the thyroid gland.

12.5 Dosimetry & Biological Effects

Each of the three types of nuclear decay, α, β, and γ, release energy in the form of the kinetic energies of the decay particle and of the recoiling nucleus. By virtue of their kinetic energies these decay products can ionize molecules in substances they pass through. Since biological molecules typically do not respond well to being randomly ionized, such ionizations constitute biological damage. The total biological damage done by a high-energy particle will be

Radiation	RBE	Radiation	Energy	RBE
X & γ rays	1	Protons	>2 MeV	5
Electrons	1	Neutrons	<10 keV	5
α particles	20	Neutrons	10 to 100 keV	10
Heavy recoil nuclei	20	Neutrons	100 keV to 2 MeV	20
Fission fragments	20	Neutrons	2 to 20 MeV	10
		Neutrons	>20 MeV	5

Table 12.1: Table of Approximate RBEs

proportional to the total number of ionizations it causes, which in turn is proportional to its energy. For this reason, tables of radionuclides usually quote decay energies along with the decay scheme and half-life. One would, for example, quote tritium as undergoing an 18.590 keV β decay with a half-life of 12.33 yr, meaning that 18.590 keV of energy are released by the decay.[13]

The first step in calculating how badly you've been fried is to determine your absorbed dose in rads, an acronym for *radiation absorbed dose*. By definition, a rad is $\frac{1}{100}\,\frac{\text{J}}{\text{kg}}$, that is, $\frac{1}{100}$ Joule of radiation energy absorbed per kilogram of the body tissue doing the absorbing. The second step is to convert this absorbed dose into an equivalent dose, a kind of "fry index" known as a rem[14] that gives a better measure of the physiological severity of your exposure by taking into account that the same absorbed dose will result in different amounts of biological damage depending on the type of radiation. The factor that converts rads into rems is known by two names: the earlier, courtesy of the bomb-makers, is RBE (*relative biological effectiveness*); those striving for political correctness may instead opt for the more sensitive modern equivalent QF (*quality factor*). While the phrase *relative biological effectiveness* reveals a very strange and disturbing perspective on radiation exposure, *quality factor* is just so wrong on so many levels that we will stick with RBE:

$$\#\text{ rem} = \#\text{ rad} \times \text{RBE}$$

The values of the RBE for various kinds of radiation are given in table (12.1).[15]

[13] Recall that an electron volt is a unit of energy, literally the electron charge times one volt (V):

$$1\text{ eV} = 1.60217653 \times 10^{-19}\text{ C} \times 1\text{ V} = 1.60217653 \times 10^{-19}\text{ J}$$

In the context of nuclear decay, one commonly deals with keV ($1\text{ keV} = 10^3\text{ eV}$) and MeV ($1\text{ MeV} = 10^6\text{ eV}$).

[14] *Rem* is an acronym for *Roentgen equivalent, man*. We won't get into the history of the *Roentgen* or *equivalent* parts, but the *man* part is the logical culmination of a series like *Roentgen equivalent, mouse*; *Roentgen equivalent, rat*; *Roentgen equivalent, wombat*; etc.

[15] These numbers are from `http://pdg.lbl.gov/2007/reviews/radiorpp.pdf`.

You will frequently also see exposures expressed in an alternative system of units, the Gray (Gy) and Sievert (Sv). Grays and Sieverts are for people who are offended by the $\frac{1}{100}$ in

$$1\,\text{rad} = \frac{1}{100}\,\frac{\text{J}}{\text{kg}}$$

and have therefore instead defined

$$1\,\text{Gy} = 1\,\frac{\text{J}}{\text{kg}} = 100\,\text{rad} \qquad\qquad 1\,\text{Sv} = 100\,\text{rem}$$

For example, suppose that as a result of a decay calculation and some very bad luck, you ascertain that you have been exposed to the decay of 2×10^{17} tritium nuclei, more or less evenly over your entire 65 kg body — what is known as whole-body exposure. Since tritium undergoes 18.590 keV β decay, your exposure in rads is

$$\frac{2 \times 10^{17}\,\text{decays}}{65\,\text{kg}} \times \frac{18.590 \times 10^{3}\,\text{eV}}{\text{decay}} \times \frac{1.6 \times 10^{-19}\,\text{J}}{\text{eV}} = 9.2\,\frac{\text{J}}{\text{kg}} \times \frac{1\,\text{rad}}{\frac{1}{100}\,\frac{\text{J}}{\text{kg}}} = 920\,\text{rad}$$

Multiplying by the RBE of 1, your exposure in rems is 920 rem — which, as you can see from the following list of symptoms and prognoses for acute exposures, means you can pretty much start the bus: [16]

0-100 rem No noticeable immediate effects, though there is always, as will be discussed below, an increased chance of subsequently developing cancer or of a genetic defect in your offspring.

100-200 rem Vomiting in 5 to 50% of victims after 3 to 6 hours. Moderate leukopenia (low white-cell count). Recovery in a matter of weeks, with hospitalization of a handful of victims for up to 60 days.

200-600 rem Vomiting in 50 to 100% of victims after 2 to 4 hours, some temporary cognitive impairment. Epilation (hair loss) over 300 rem. Hematopoietic (blood-producing) and respiratory systems principally affected. Severe leukopenia, hemorrhaging, purpura (you don't want to know). Low mortality with medical treatment, but hospitalization for 60 to 90 days required for 90% of victims. Critical period 4 to 6 weeks.

600-800 rem Vomiting in 75 to 100% of victims after 1 to 2 hours, cognitive impairment. Hematopoietic and respiratory systems principally affected. Severe leukopenia, purpura, hemorrhaging, epilation. High mortality even with medical treatment, hospitalization for more than 90 days. Critical period 4 to 6 weeks.

[16] This information paraphrased from the Department of Veteran Affairs publication *Terrorism with Ionizing Radiation General Guidance Pocket Guide* (`http://www.afrri.usuhs.mil/www/outreach/pdf/pcktcard.pdf`). How bizarre is that?

800-3000 rem Onset of symptoms in less than an hour. Incapacitation. Vomiting, diarrhea, fever, electrolyte disturbance. Damage to gastrointestinal tract. Critical period 2 to 14 days, but you are almost certainly toast.

Over 3000 rem Onset of symptoms in minutes. Incapacitation. Convulsions, ataxia (you lose your coordination), tremor, lethargy. Damage to gastrointestinal tract. Critical period 1 to 48 hours. You are burnt toast.

At a microscopic level, radiation does biological damage by ionizing biological molecules: electrons are ripped out of the molecules by the electrical attraction or repulsion from passing α and β particles, or by absorbing some of the energy and momentum of a passing γ ray.[17] Understandably, biological molecules generally do not respond well to this kind of arbitrary ionization. Most of the time the molecules simply cease to function and become cellular flotsam that your body disposes of by the same mechanism it uses to dispose of other worn out or damaged molecules. In really massive acute exposures, radiation damages so many molecules at once that your body's clean-up and repair mechanisms are overwhelmed, with the result that you die just as you would from other sorts of massive internal injuries. With less intense exposures, your body is able to clean up and repair most of the damage, but there is a small chance that the radiation will have damaged a molecule in just the right way to cause cancer or, if the molecule is DNA in a reproductive cell, to cause a genetic defect in your offspring.[18] There is no doubt that radiation can cause cancer and genetic defects, but determining the quantitative correlation between exposure in rems and incidences of cancers and genetic defects is very tricky. If one studies the long-term effects of small exposures on large populations, it is very difficult to statistically separate the effects of radiation from those of the myriad other variables like diet, etc. On the other hand, while it is possible to get good, clear data from studies of people who received known massive acute exposures, it is not clear that the results

[17] Neutron radiation — the loose neutrons produced by fission and fusion processes — do not have electric charge and therefore do not cause ionization. But exposure to neutrons is still harmful because they give rise to secondary radiation: when neutrons are absorbed by otherwise innocuous nuclei, some of the isotopes produced are radioactive, and neutrons can also knock nuclei into excited states that subsequently undergo gamma decay. This was, in fact, the idea behind the neutron bomb, a kind of nuclear weapon designed to maximize neutron flux: the neutrons would penetrate tanks and other heavy shielding far more easily than other forms of radiation, and the resulting secondary radiation would be lethal.

[18] Actually, it would be more accurate to say *random genetic mutation*; the change could be beneficial, it's just that it's far more likely not to be.

of these studies can be extrapolated to the lower, longer-term exposures experienced by the general population. Under the linear hypothesis or linear no-threshold model it is assumed that the carcinogenic and genetic effects of radiation *per rem* are the same for massive acute and for lower, longer-term exposures, but it could very well be that your body is better — or, conceivably, not as well — able to repair the damage done by milder exposures over long periods of time.

The more deadly a particular kind of radiation, that is, the higher its RBE, the easier it is to shield yourself from it: the tendency of the radiation to lose energy by ionizing molecules is the same for the molecules of any other substance as it is for the molecules of your body.[19] Alpha particles — big, blundering, lumbering particles of +2 electric charge — lose their energy very quickly in any kind of matter, so that even thin clothing is enough to stop most of them. Beta particles can penetrate farther but are still easily stopped by just a few millimeters of reasonably dense material. Gamma rays, the least deadly but also the hardest to stop, may, depending on the intensity of the source, require several centimeters or decimeters of shielding made of lead or some other very dense substance with heavy atoms.

Throughout your life, you are continually exposed to radiation from a variety of natural sources. This exposure has very little to do with man-made sources like fallout from weapons tests or reactor accidents; it is the same for you as it was for your ancestors eons ago, and in the United States ranges from about 40 to 400 mrem per year, with an average of about 360 mrem.[20] The sources of this natural background exposure are:

- About 200 mrem/yr from radon.[21,22] Radon is one of the radionuclides produced in the series of decays that starts from uranium. It is a noble gas and, once produced by the decay of its immediate parent, radium, it leeches up through the soil into the air or is brought up with recently pumped ground water. Depending on the geology where you live and on factors like the porosity of your basement and how well ventilated your home is, your exposure from radon can vary by two orders of magnitude or more. As will be discussed in the next section, radon exposure is a significant cause of lung cancer.

[19] As noted in footnote 17, neutrons are an exception to this. Shielding against neutrons requires substances dense in hydrogen and other light nuclei that sap the energy of the neutrons in collisions and then absorb them.

[20] `http://pdg.lbl.gov/2007/reviews/radiorpp.pdf`.

[21] Ibid.

[22] We probably should note that this component of your background exposure was not the same for your ancestors eons ago, because they didn't live in energy-efficient homes that sealed radon in instead of allowing it to disperse into the open atmosphere.

- Roughly equal thirds of the rest of your background exposure come from
  - The Earth around you: The Earth coalesced out of dust from exploded stars, some of which consisted of radionuclides with extremely long half-lives, such as isotopes of uranium. These elements have been slowly decaying away since the Earth formed and will continue to do so for eons to come.
  - Outer space: There is a lot of violent stuff happening out there among the stars of the universe, radiation from some of which happens to fall on the Earth. Much of this is filtered out by the atmosphere, but some makes it through to ground level.
  - Your own body: Physically, at least, people are basically dirt bags:[23] Dirt is just broken-up rock, some of which plants absorb through their roots. You in turn eat plants, or animals that fed on plants, or animals that fed on animals that fed on plants, ..., with the result that you are in part composed of the same material as the dirt around you, which is somewhat radioactive.

    Those of you who like bananas will be happy to learn that the principal radionuclide in your body is potassium-40.

A few other figures to put radiation exposure in perspective:

- In the United States, regulations limit whole-body exposures of those who work in the nuclear industry to 5000 mrem/yr and of the general public to 100 mrem/yr.[24]
- You get about 1 mrem of exposure from an ordinary dental x-ray, 6 mrem from a chest x-ray.[25]
- In Colorado, Wyoming, New Mexico, and Utah, background exposure is about twice the national average, partly because of the high uranium content of the mountain soil and partly because there is less atmospheric filtering of cosmic radiation due to the high altitude.[26] In Kerala and Madras, India, background exposures average 1500 mrem per year from non-radon sources, with a comparable additional exposure from radon.[27]

[23] We suppose if you want to go back a generation further, you could also argue that people are basically star dust.

[24] http://www.nrc.gov/reading-rm/doc-collections/cfr/part020/full-text.html

[25] Bernard L. Cohen, *The Nuclear Energy Option*, available on-line. In particular, the exposure information we present here may be found in Chapter 5: `http://www.phyast.pitt.edu/~blc/book/chapter5.html`.

[26] Ibid.

[27] `http://www.doh.wa.gov/ehp/rp/factsheets/factsheets-htm/fs2rad&life.htm`.

- Bricks are made of earth: living in a brick rather than a wooden house can raise your background exposure by about 10%.[28]

- Statistically separating cancers due to radiation exposure from cancers due to other causes is extremely difficult to do for longterm, low-level exposures, but under the linear hypothesis discussed on p.652 estimates are typically on the order of 10^{-4} cancers per rem. Now, either you are the one who gets cancer or you aren't, but if the estimate is, say, 2×10^{-4} cancers per rem, then out of 10,000 people each exposed to 1 rem of radiation, on average two people would eventually develop a cancer as a result of their exposure.

- On average, each millirem of exposure corresponds to 2 min of decreased life expectancy.[29]

- In terms of carcinogenic potential, one cigarette is equivalent to about 5 mrem.[30]

- Estimates of the rates at which radiation-induced genetic defects are passed along to offspring also vary, but are typically on the order of 10^{-5} genetic defects per rem.

- In terms of potential to cause genetic defects, 1 mrem is equivalent to about 5 hr of wearing pants.[31]

In a 1991 report, *Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants* (NUREG-1150),[32] the Nuclear Regulatory Commission analyzes the probabilities and consequences of catastrophic events at

[28]Cohen, op. cit.

[29]Ibid.

[30]Ibid.

[31]In case it's not immediately obvious how wearing pants can cause genetic defects: sperm, it turns out, are very sensitive to heat, and wearing pants holds the heat in. So if you're a guy you may want to think about switching to kilts.

This very rough estimate of equivalence apparently comes from Ehrenberg, von Ehrenstein, & Hedgram, "Gonad Temperature and Spontaneous Mutation Rate in Man," *Nature*, December 2, 1433 (1957), though we must confess that we've only ever seen it quoted in secondary sources because we're far too lazy to dig out the original source.

[32]If you have a fast enough connection that a 42 Mb PDF file doesn't scare you and have the stomach for wading through 691 pages of government-reportese, you can download the report for yourself at `http://www.nrc.gov/reading-rm/adams/web-based.html` by using the search link to find `ML040140729`. NUREG-1150 supersedes an earlier, less sophisticated, and more pessimistic analysis known as WASH-1400 or the Rasmussen report (*An Assessment of Accident Risks in US Commercial Nuclear Power Plants — Calculation of Reactor Accident Consequences*), which you can download from the same site by searching for `ML053290245`.

five fission reactors chosen to be representative of the variety of reactors in use in the United States. According to this report:[33]

- An accident resulting in damage to the core, that is, a partial or total meltdown, is expected to occur on average once every 17000 to 250,000 reactor-years (ry) of operation.[34] This estimate is for internal events; the probability due to external events such as earthquakes can be significantly higher. Note, however, that core damage does not necessarily entail a breach of containment that would release radionuclides.

- An "early failure or bypass" of containment is expected to occur on average once every 200,000 to 1,000,000 ry.[35]

- An accident resulting in zero fatalities from acute exposure but up to 700 eventual cancer deaths can be expected once every 100,000 ry. The total exposure of the population would be up to 3,000,000 rem (up to 100,000 rem within a 50-mile radius).[36]

- An accident resulting in up to 4000 fatalities from acute exposure and up to 100,000 eventual cancer deaths can be expected once every 10^9 ry. The total exposure of the population would be up to 4×10^8 rem.[37]

- The average number of fatalities from acute exposure is estimated to be from 9×10^{-9} to 8×10^{-4} per reactor-year and the number of eventual cancer deaths from 1×10^{-3} to 3×10^{-2} per reactor-year.[38]

- A large release (defined as a release resulting in at least one death from acute exposure) can be expected once every 2,000,000 ry.[39]

There are no quick fixes or easy answers to the problem of generating power in a safe and environmentally sustainable way. Even if stringent conservation measures were universally implemented, we would still need to produce vast amounts of energy, and we are not likely to meet those needs solely by wind, solar, geothermal, hydroelectric, and other environmentally friendly means — at least not in the near future, and perhaps not ever.

Nuclear fusion would be safe and nonpolluting, and the supply of fuel is virtually limitless. Research into this extremely promising technology has of course therefore been woefully underfunded. But even with funding it would

[33] At least, within the limits of the ability of our rather weak eyes to read the irritatingly small logarithmic plots in the report.

[34] NUREG-1150, p.8–2.

[35] Ibid., p.9–6.

[36] Ibid., p.11–3f.

[37] Ibid., p.11–3f.

[38] Ibid., p.12–2.

[39] Ibid., p.13–10.

likely take years of research for fusion to become a viable means of power generation.

Fission technology is here now, but in addition to the risk of an accident there is the certain problem of disposing of the highly radioactive spent fuel. There is still no longterm solution to that problem even though reactor sites are running out of space to store the waste already produced. We need to find a location and storage medium that will remain safe from corrosion and from earthquakes and other geological events for the eons it will take the levels of radioactivity to subside to reasonably safe levels. At present the only site being seriously considered is Yucca Mountain, based on the irrefutably sound scientific observation that there are very few voters in Nevada.[40]

And as if to rub it in that you just can't win, conventional power plants that burn fossil fuels not only contribute to global warming and acid rain, but also release significant radioactivity into the atmosphere: coal and oil, like everything else around us, contain small amounts of naturally occurring radionuclides, which are carried up and out of the smokestacks together with the combustion products.

Like exposure from natural sources, exposure to radioactivity from power production is therefore not a yes-no question, but a question of degree, of how much exposure and risk are acceptable. These are questions you will have to decide, and you can't make a wise decision without some understanding of radioactivity and the effects of exposure to it.

12.6 One More Reason New Jersey Is a Disgusting Place

New Jersey isn't famous just for swamp gas, organized crime, high taxes, pervasive corruption, illegally dumped medical waste, nanny laws, high auto-insurance rates, Superfund clean-up sites, and a general lack of culture and manners — it's also a great place to expose yourself to radon.[41]

This exposure from radon begins with radium-226 in the ground: one of the rungs in a ladder of decays that starts from uranium, radium-226 α-

[40] At this point someone always seems to suggest launching our nuclear waste into space. While the Sun would be happy to gobble up all our waste, there is far more of it than we could, as a practical matter, ever hope to launch. And even if we could launch it all, the problem would be getting from here to there: mishaps with rockets are not uncommon, and we would then be faced with the possibility of massive contamination of the atmosphere, oceans, or land by an explosion or crash.

[41] `http://www.epa.gov/radon/zonemap/newjersey.htm` maps out radon levels in New Jersey, and you can find Superfund lists on other EPA sites. But the EPA is curiously silent about the swamp gas, organized crime, high taxes, pervasive corruption, illegally dumped medical waste, nanny laws, high auto-insurance rates, and general lack of culture and manners.

decays into radon-222, a noble gas that seeps into your house through the basement or is brought up with ground water. If a radon nucleus happens to decay while inside your lungs, it turns into polonium-218, which in turn is radioactive and gives rise to a series of α and β decays that eventually end in a stable isotope of lead. So every radon nucleus that decays in your lungs ends up zapping your lung tissue several times over. Which is bad news if you're worried about lung cancer.

While tighter sealing of the basement can reduce the seepage of radon into your home, the simplest, cheapest, and most effective way to reduce exposure is to keep your home well ventilated, so that you are breathing more fresh air and less radon.[42] Unfortunately, better ventilation means less efficient heating and cooling — so that you have to pay higher fuel bills to have the power company supply more power by burning more fossil fuel that releases more radioactive isotopes and chemical pollution into the atmosphere and ends up poisoning the air that you breathe anyway. Another fine illustration of the principle that you just can't win. Ever. No matter what.

The only real solution, of course, is to move out of New Jersey.

[42]To the extent that the air in New Jersey can be considered fresh. But after you've been here a while your nose gets used to the swamp gas and landfills.

12.7 Problems

1. Write the following decay schemes, showing all products, after the fashion of §12.2. If you need it, there is a periodic table on p.663.

   (a) Radium-226 α-decays into radon.

   (b) This radon subsequently α-decays into polonium.

   (c) Tritium (hydrogen-3) undergoes β decay.

   (d) An excited state of lead-207 undergoes γ decay.

   (e) Iodine-131 undergoes simultaneous β and γ decay.[43]

2. Determine the energy released in the fusion reaction

$$^{2}_{1}\mathrm{H} + {}^{3}_{1}\mathrm{H} \rightarrow {}^{4}_{2}\mathrm{He} + {}^{1}_{0}\mathrm{n}$$

   given

$^{2}_{1}$H	2.01410178 u	$^{3}_{1}$H	3.01602931 u
$^{4}_{2}$He	4.00260325 u	$^{1}_{0}$n	1.00866492 u

   and, on the presumption that you're too **** lazy to look them up yourself, $c = 2.99792458 \times 10^{8}$ m/s and 1 u $= 1.66053886 \times 10^{-27}$ kg.

3. You plan to market a new sports drink called *Zap*, which consists solely of water in which one of the two hydrogens in each water molecule (H_2O) will be tritium ($^{3}_{1}$H). Tritium undergoes 18.590 keV β decay with a half-life of 12.33 yr and an RBE of 1.

   (a) If this were ordinary water, a one-liter bottle would have a mass of 1000 g. What is its mass when one of the two hydrogens in each H_2O molecule is tritium?

   (b) How many nuclei of tritium are in this sample?

   (c) How many of the tritium nuclei decay in the first six months?

   (d) What is the initial activity of the sample in Curies?

   (e) What is the activity of the sample in Curies after six months?

   (f) How long will it take for 99% of the sample to decay?

   (g) How many of the tritium nuclei decay in the first minute?

   (h) If you are 65 kg and have chugged down a full bottle of *Zap*, with how many rem per minute are you being zapped?

   (i) Is all this arithmetic really tedious or what?

[43] Many nuclides undergo such decays, although calling them simultaneous is a bit of a misnomer: the decays are distinct, but occur in such rapid succession that they are effectively simultaneous.

4. Suppose you have a sample of pure ^{226}Ra, which decays with a half-life of 1602 yr. How many grams would the sample have to be in order for the activity to be 1.0 Ci?

5. In the spirit of daring scientific inquiry, you create a new element by putting a fork in the microwave and nuking it on high for ten minutes.[44] Using the mass spectrometer you "borrowed" from the science building, you isolate 12 mg of this new element, which proves to be radioactive. After 3.0 hr, the activity of the sample is only 40% of what it was initially.

   (a) What is the half-life of this new element?

   (b) How much longer will it be until the activity drops to 16% of what it was initially?

6. As a result of bombardment with cosmic radiation from outer space, small amounts of carbon-14 are continually being formed at a constant rate. This $^{14}_{6}$C undergoes β decay with a half-life of 5700 yr, and as a result of the equilibrium between its rates of production and decay the proportion of $^{14}_{6}$C in the atmosphere and oceans has remained constant through the ages.[45] Since living plants and animals are continually ingesting carbon, they contain this same proportion of $^{14}_{6}$C. You have, however, probably noticed that most things that eat and breathe cease eating and breathing when they die. One happy consequence of this is that there's more food to eat and air to breathe for those of us who are still kicking. Another happy consequence is that the decay of $^{14}_{6}$C in the tissues of dead plants and animals makes it possible for us to date their demise: we simply compare the proportion of $^{14}_{6}$C remaining in the plant or animal artifact to the proportion of $^{14}_{6}$C in currently living things. By this means organic material can be quite accurately dated as far back as 60,000 yr or so. (In older artifacts too little $^{14}_{6}$C remains for accurate measurement.)

   Suppose, for example, that archaeologists excavating an Ancient Egyptian school find a wad of gum under a desk and, upon analysis, find that it has only 38% as much $^{14}_{6}$C as a fresh gum-wad of identical composition. How old is the gum-wad?

[44] In reality microwaves are of course far below the frequencies and energies needed to induce nuclear transitions. For those whose otherwise irresistible curiosity would lead them to become candidates for a Darwin Award: the actual result would be a lot of sparking, the destruction of the fork and the microwave oven, and possibly even a fire.

[45] Actually, this isn't quite true; there are slight variations with time and location that have to be taken into account in very precise work.

7. The dining hall buys some apples grown locally on a Superfund clean-up site contaminated with ${}^{30}_{13}$Nj (New Jersium),[46] which undergoes 100 keV β decay with a half-life of 8 hr and an RBE of 1. You eat an apple containing, at the moment you ingest it, 2 μg of ${}^{30}_{13}$Nj. Assuming (very unrealistically) that your exposure is evenly spread over your entire 70 kg body, how long will it take you to receive an almost certainly fatal dose of 800 rem? (The numbers are such that you should be able to do the arithmetic without a calculator.)

8. Alexander Litvinenko was poisoned with polonium-210, which undergoes 5.30438 MeV α decay with a half-life of 138.376 day and an RBE of 20. Under the assumptions that he was 80 kg, that he ingested 50 mCi, and that his exposure was uniform over his whole body, how many rem did he receive on the first day?

9. Doses of Iodine-131 with activities of 4 to 10 mCi are typically administered in radiotherapy of hyperthyroidism. ^{131}I undergoes simultaneous β and γ decay with a half-life 8.0197 day, typical energies of 606 and 364 keV, respectively, and an RBE of 1.

   (a) How many grams of ^{131}I correspond to an activity of 10 mCi?

   (b) Under the somewhat unrealistic assumption that all of the 10 mCi of radiation is absorbed by a 20 g thyroid gland, what total dose in rem will the gland ultimately receive?

10. About 17% of people ultimately die of cancers due to various causes. If you receive the average background dose of 360 mrem/yr throughout a 70 yr lifetime and the incidence is 8×10^{-4} fatal cancers per rem (a pretty liberal estimate), what is your chance of developing a cancer attributable to background radiation exposure?

[46] You won't find New Jersium in the periodic table. It smells so bad that Mendeleev left it out.

11. Find and analyze a web site or an article from a newspaper or magazine related to radiation and its effect on people. The site or article should cite enough numerical data for you to do some significant quantitative analysis and thus make some assessment the validity or reasonableness of its content. Commercial, amateur, and political sites are often interesting — be skeptical of their assertions. You might look into atmospheric weapons testing and fallout, radon, power production by nuclear fission and fusion, Chernobyl, Three-Mile Island, nuclear generators in satellites, or disposal of nuclear waste from weapons and reactors; there is a ton of stuff out there — sensible and otherwise.

    You should not attempt to find a great breadth of material; you just need a few good numbers to work with. You will then do your analysis, write up your conclusions, and submit them along with a printout of the relevant parts of the original web page or article. Your should show your calculations explicitly, accompanied by enough explanation to make them clear.

    If you are using a search engine to find a web site, doing an "advanced search" with keywords like `rem`, `mrem`, `curies`, `ci`, or the like will increase your chances of finding a site with usable numerical data.

12.8 Sketchy Answers

(2) $2.81513595 \times 10^{-12}$ J.

(3a) 1111 g.

(3b) 3.346×10^{25}.

(3c) 9.273×10^{23}.

(3d) 1.610×10^{6} Ci.

(3e) 1.566×10^{6} Ci.

(3f) 81.92 yr.

(3g) 3.576×10^{18}.

(3h) 16390 rem.

(4) *Surprise!* 1.0 g.

(5a) 2.3 hr.

(5b) Ha, ha: 3.0 hr.

(6) 8000 yr.

(7) 24 hr.

(8) 3400 rem.

(9a) 8.0×10^{-8} g.

(9b) 2.8×10^{5} rem.

(10) 2%.

Table 4.1. Revised 2005 by C.G. Wohl (LBNL). Adapted from the Commission on Atomic Weights and Isotopic Abundances, "Atomic Weights of the Elements 1999," Pure and Applied Chemistry **73**, 667 (2001), and G. Audi, A.H. Wapstra, and C. Thibault, Nucl. Phys. **A729**, 337 (2003). The atomic number (top left) is the number of protons in the nucleus. The atomic mass (bottom) is weighted by isotopic abundances in the Earth's surface. Atomic masses are relative to the mass of the carbon-12 isotope, defined to be exactly 12 unified atomic mass units (u). Errors range from 1 to 9 in the last digit quoted. Relative isotopic abundances often vary considerably, both in natural and commercial samples. A number in parentheses is the mass of the longest-lived isotope of that element—no stable isotope exists. However, although Th, Pa, and U have no stable isotopes, they do have characteristic terrestrial compositions, and meaningful weighted masses can be given. For elements 110 and 111, the numbers of nucleons A of confirmed isotopes are given.

PERIODIC TABLE OF THE ELEMENTS

1 IA												13 IIIA	14 IVA	15 VA	16 VIA	17 VIIA	18 VIIIA
1 H Hydrogen 1.00794	2 IIA																2 He Helium 4.002602
3 Li Lithium 6.941	4 Be Beryllium 9.012182											5 B Boron 10.811	6 C Carbon 12.0107	7 N Nitrogen 14.0067	8 O Oxygen 15.9994	9 F Fluorine 18.9984032	10 Ne Neon 20.1797
11 Na Sodium 22.989770	12 Mg Magnesium 24.3050	3 IIIB	4 IVB	5 VB	6 VIB	7 VIIB	8	9 VIII	10	11 IB	12 IIB	13 Al Aluminum 26.981538	14 Si Silicon 28.0855	15 P Phosph. 30.973761	16 S Sulfur 32.065	17 Cl Chlorine 35.453	18 Ar Argon 39.948
19 K Potassium 39.0983	20 Ca Calcium 40.078	21 Sc Scandium 44.955910	22 Ti Titanium 47.867	23 V Vanadium 50.9415	24 Cr Chromium 51.9961	25 Mn Manganese 54.938049	26 Fe Iron 55.845	27 Co Cobalt 58.933200	28 Ni Nickel 58.6934	29 Cu Copper 63.546	30 Zn Zinc 65.39	31 Ga Gallium 69.723	32 Ge German. 72.64	33 As Arsenic 74.92160	34 Se Selenium 78.96	35 Br Bromine 79.904	36 Kr Krypton 83.80
37 Rb Rubidium 85.4678	38 Sr Strontium 87.62	39 Y Yttrium 88.90585	40 Zr Zirconium 91.224	41 Nb Niobium 92.90638	42 Mo Molybd. 95.94	43 Tc Technet. (97.907216)	44 Ru Ruthen. 101.07	45 Rh Rhodium 102.90550	46 Pd Palladium 106.42	47 Ag Silver 107.8682	48 Cd Cadmium 112.411	49 In Indium 114.818	50 Sn Tin 118.710	51 Sb Antimony 121.760	52 Te Tellurium 127.60	53 I Iodine 126.90447	54 Xe Xenon 131.293
55 Cs Cesium 132.90545	56 Ba Barium 137.327	57–71 Lantha-nides	72 Hf Hafnium 178.49	73 Ta Tantalum 180.9479	74 W Tungsten 183.84	75 Re Rhenium 186.207	76 Os Osmium 190.23	77 Ir Iridium 192.217	78 Pt Platinum 195.078	79 Au Gold 196.96655	80 Hg Mercury 200.59	81 Tl Thallium 204.3833	82 Pb Lead 207.2	83 Bi Bismuth 208.98038	84 Po Polonium (208.982430)	85 At Astatine (209.987148)	86 Rn Radon (222.017578)
87 Fr Francium (223.019736)	88 Ra Radium (226.025410)	89–103 Actinides	104 Rf Rutherford. (261.10877)	105 Db Dubnium (262.1141)	106 Sg Seaborg. (263.1221)	107 Bh Bohrium (262.1246)	108 Hs Hassium (277.1498)	109 Mt Meitner. (268.1387)	110 Ds Darmstadt. [269,271]	111 [272]							

Lanthanide series	57 La Lanthan. 138.9055	58 Ce Cerium 140.116	59 Pr Praseodym. 140.90765	60 Nd Neodym. 144.24	61 Pm Prometh. (144.912749)	62 Sm Samarium 150.36	63 Eu Europium 151.964	64 Gd Gadolin. 157.25	65 Tb Terbium 158.92534	66 Dy Dyspros. 162.50	67 Ho Holmium 164.93032	68 Er Erbium 167.259	69 Tm Thulium 168.93421	70 Yb Ytterbium 173.04	71 Lu Lutetium 174.967
Actinide series	89 Ac Actinium (227.027752)	90 Th Thorium 232.038055	91 Pa Protactin. 231.035884	92 U Uranium 238.02891	93 Np Neptunium (237.048173)	94 Pu Plutonium (244.064204)	95 Am Americ. (243.061381)	96 Cm Curium (247.070354)	97 Bk Berkelium (247.070307)	98 Cf Californ. (251.079587)	99 Es Einstein. (252.08298)	100 Fm Fermium (257.095105)	101 Md Mendelev. (258.098431)	102 No Nobelium (259.1010)	103 Lr Lawrenc. (262.1096)

Chapter 13

Thermal Physics

13.1 Statistical Mechanics Versus Thermodynamics

Up to now, aside from a few parenthetic observations about what happens to the energy apparently lost to friction, we have dealt with energy only at the macroscopic level, that is, at the level of objects of "everyday" sizes — as opposed to the microscopic level of atoms and molecules. Macroscopically, energy was of two types: the kinetic energy due to the translational and rotational motions of bodies, and the various potential energies corresponding to the forces that bodies exert on each other. This same dichotomy of energy of course holds microscopically as well: there is the kinetic energy due to the motions of molecules and the various potential energies corresponding to the forces that atoms and molecules exert on each other or to external forces. In principle, there is therefore no reason why we couldn't study microscopic motion by the same methods of dynamics and conservation of energy that we used to study macroscopic motion. The problem with this approach, however, is that macroscopic samples of substances typically involve on the order of Avogadro's number of molecules, and as a practical matter we cannot keep track of and do calculations for systems of so many bodies.

To get around this problem, thermal physics deals, not with the motions of individual molecules, but with the average values of physical quantities over large numbers of molecules. Even though the dynamics of the molecules are completely deterministic and could in principle be calculated exactly,[1] they are treated statistically in thermal physics. We can get away with this "laziness" because the statistical fluctuations in a physical quantity averaged over N molecules can be shown to be of relative size $1/\sqrt{N}$. For macroscopic

[1] At least in Newtonian physics. But in our study of things thermal, we aren't going to be worried about quantum theory and the truly statistical nature of reality; a discussion of that will have to wait until Chapter 22.

samples with on the order of Avogadro's number of molecules, the statistical fluctuations are typically one part in $\sqrt{6 \times 10^{23}} \approx 10^{12}$, and the average values calculated by statistical methods are therefore exceedingly accurate for such systems.

One approach to the statistical analysis of thermal systems is thermodynamics: the states of systems are expressed in terms of relations among the aggregate or average values of physical quantities, and laws governing changes of state are discerned from empirical observation. For example, the ideal-gas law $pV = nRT$, with which you are no doubt familiar from a previous life in chemistry, expresses the state of a sample of n moles of gas as a relation among its average pressure p, its average temperature T, and the aggregate volume V it occupies.[2] And, as we will see later, changes in the state of this sample of gas must obey a thermal version of conservation of energy that takes into account that heat is a form of energy.

Thermodynamics was the way in which thermal physics was first formulated, and like any product of blundering around in the dark is rather homely and inelegant. As you will discover if you continue your studies of physics, modern thermodynamics consists chiefly of playing games with relations obtained by taking partial derivatives of some quantity W with respect to some variable X while holding variables Y, Z, ... constant, then with respect to Y while holding X, Z, ... constant, and so on.

The more fundamental and insightful approach to thermal statistical analysis is statistical mechanics. The fundamental tenet of statistical mechanics is that the probability of a molecule (or system of molecules) being in a state of energy E is proportional to $e^{-E/kT}$, where T is the temperature and $k = 1.3806505 \times 10^{-23}$ J/K is the Boltzmann constant, a conversion factor between the artificial units of degrees in which temperature is conventionally measured and the energy units in which it should properly be measured.[3, 4, 5] From this one simple tenet it is possible to derive all of the results of thermodynamics and much more besides.

The logic behind the factor $e^{-E/kT}$, which is known as the Boltzmann factor, is probably not immediately apparent, but then, Boltzmann wouldn't have gotten much credit for it if it were obvious, now would he?[6] Anyway,

[2] For those who like jargon, quantities that, like pressure or temperature, have a value at each point are called intensive; aggregate quantities like volume are called extensive.

[3] Actually, one can take several different approaches to statistical mechanics, each starting from its own vantage point. It is also possible to derive the $e^{-E/kT}$ from the alternative tenet that a system is equally likely to be in any of the multiplicity of microscopic states that give rise to the same macroscopic state.

[4] Ludwig Boltzmann was, along with James Clerk Maxwell, one of the principal architects of statistical mechanics.

[5] Note that, unlike the speed of light c and Plank's constant h, the Boltzmann constant k is therefore *not* a fundamental physical constant.

[6] Remember, this was back in the days when the blatantly self-evident ideas like one-click shopping were not as highly esteemed as they are today.

we can make the Boltzmann factor plausible by recalling that a potential energy U can be shifted by an arbitrary additive constant without affecting the physics, that is, that what is physically meaningful is the *change* in potential energy as we go from one location to another. Since the total energy $E = K+U$ involves this same arbitrary additive constant, a physically meaningful result such as the relative likelihood of a system being in states of energy E' and E should depend only on the energy difference $E' - E$ between the states. Now, the probability $p(E')$ of being in a state of energy E' relative to the probability $p(E)$ of being in a state of energy E is the ratio of the probabilities,[7] $p(E')/p(E)$, and this relative probability should be some function f of $E' - E$:

$$\frac{p(E')}{p(E)} = f(E' - E) \tag{13.1}$$

When $E' = E$, eq. (13.1) reduces to

$$\frac{p(E)}{p(E)} = f(0)$$

so that

$$f(0) = 1 \tag{13.2}$$

To determine $p(E)$, we need a relation in terms solely of $p(E)$, E, and known quantities, and we can generate such an equation from eqq. (13.1) and (13.2) by looking at the case $E' = E + dE$: to first order in the infinitesimal dE, we have

$$p(E') = p(E) + dE\,\frac{dp(E)}{dE}$$

and, if we denote the derivative of f by f',

$$\begin{aligned} f(E' - E) &= f(dE) \\ &= f(0) + dE\,f'(0) \\ &= 1 + dE\,f'(0) \end{aligned}$$

Using these results in eq. (13.1), we then have

$$\begin{aligned} \frac{p(E) + dE\,\frac{dp(E)}{dE}}{p(E)} &= 1 + dE\,f'(0) \\ \frac{p(E) + dp(E)}{p(E)} &= 1 + dE\,f'(0) \end{aligned}$$

[7] As in: if there is a 60% chance of X and a 30% chance of Y, then the probability of X relative to Y is $60/30 = 2$, that is, X is twice as likely as Y.

$$\begin{aligned} 1 + \frac{dp(E)}{p(E)} &= 1 + dE\, f'(0) \\ \frac{dp(E)}{p(E)} &= dE\, f'(0) \\ \int \frac{dp(E)}{p(E)} &= \int f'(0)\, dE \\ \ln p(E) &= f'(0)\, E + \text{const} \end{aligned}$$

where we have noted that $f'(0)$, although its value remains unknown, is just a constant. Solving for $p(E)$ we thus arrive at

$$p(E) = (\text{const})\, e^{f'(0)\, E}$$

or, if we write the constants involved a bit more æsthetically as $\mathcal{N}$ and $-\beta$,

$$p(E) = \mathcal{N} e^{-\beta E} \tag{13.3}$$

The final step in establishing that $p(E)$ is proportional to $E^{-E/kT}$ is to show that $\beta = 1/kT$, which could be done by using the probability $p(E)$ to calculate a quantity for which a thermodynamic relation is already known and comparing the two results. Alternatively, we could take $\beta = 1/kT$ to define temperature, with k, as noted above, being a conversion factor between the artificial units of degrees in which temperature is conventionally measured and the energy units in which it should properly be measured. We will take the easy way out and do the latter.

The other constant in eq. (13.3), $\mathcal{N}$, is then determined by the condition that the sum of the probabilities $p(E)$ over all possible energies E must be unity, that is, 100%:

$$\int dE\, \mathcal{N} e^{-\beta E} = 1$$

Mathematically, factors like $\mathcal{N}$ are known as normalization factors.

In the next section we will show you how you apply eq. (13.3). But to summarize before we leap into the fray:

- Thermal physics resorts to statistical methods only because we cannot, as a practical matter, deal with the huge numbers of molecules in macroscopic samples.
- These statistical methods yield relations among average and aggregate values of physical quantities.
- There are statistical fluctuations in these average values. For macroscopic bodies these fluctuations are, however, exceedingly small, and the averages are therefore exceedingly accurate.

13.2 Some (*ugh!*) Chemistry

Suppose you are dealing with a chemical reaction like

$$a\mathrm{A} + b\mathrm{B} \leftrightharpoons c\mathrm{C} + d\mathrm{D}$$

where a, b, c, and d are the stoichiometric coefficients of chemical species A, B, C, and D. In chemistry courses you were probably told that the concentrations $[A]$, $[B]$, $[C]$, $[D]$ of these species at equilibrium were related by

$$K = \frac{[C]^c\,[D]^d}{[A]^a\,[B]^b} \tag{13.4}$$

where K is called the equilibrium constant of the reaction. Sometimes eq. (13.4) is even called the "law of mass action," as though this name made any sense, and as though this relation were some sort of fundamental physical law.

Let us instead consider the above reaction in a sensible way. If ϵ_a, ϵ_b, ϵ_c, ϵ_d represent the molecular binding energies of A, B, C, D — that is, the electromagnetic potential energies associated with the atoms of these species being bound together — , then the total energy when you have $a\mathrm{A} + b\mathrm{B}$ is $a\epsilon_a + b\epsilon_b$, and $c\epsilon_c + d\epsilon_d$ when you have $c\mathrm{C} + d\mathrm{D}$. According to the basic axiom of statistical mechanics, the probability of having $c\mathrm{C} + d\mathrm{D}$ relative to the probability of having $a\mathrm{A} + b\mathrm{B}$ is therefore

$$\frac{e^{-(c\epsilon_c + d\epsilon_d)/kT}}{e^{-(a\epsilon_a + b\epsilon_b)/kT}} = \frac{\left(e^{-\epsilon_c/kT}\right)^c \left(e^{-\epsilon_d/kT}\right)^d}{\left(e^{-\epsilon_a/kT}\right)^a \left(e^{-\epsilon_b/kT}\right)^b} \tag{13.5}$$

Now, the concentrations of species A, B, C, and D will be proportional to their probabilities of occurrence, which in turn are proportional to $e^{-\epsilon_a/kT}$, $e^{-\epsilon_b/kT}$, $e^{-\epsilon_c/kT}$, and $e^{-\epsilon_d/kT}$. The right-hand side of eq. (13.5) can therefore be rewritten in terms of concentrations as

$$\frac{[C]^c\,[D]^d}{[A]^a\,[B]^b}$$

which is exactly the expression for the equilibrium constant.

From this we see that

- The "law of mass action" is simply a direct consequence of statistical mechanics.
- The equilibrium constant K is not a magic number; it is determined by the binding energies of the various chemical species involved in the reaction.

- The equilibrium constant K is not constant: it depends on temperature.[8] If we denote the total energy of the products and reactants by $\epsilon_p = c\epsilon_c + d\epsilon_d$ and $\epsilon_r = a\epsilon_a + b\epsilon_b$, respectively, then

$$K = \frac{e^{-\epsilon_p/kT}}{e^{-\epsilon_r/kT}} = e^{-\Delta\epsilon/kT}$$

  where $\Delta\epsilon = \epsilon_p - \epsilon_r$ is the energy difference between the aggregate products and the aggregate reactants. Thus the lower the temperature, the larger the absolute value of the exponent and the more lopsided K: at lower temperatures, the state (reactant or product) with the lower energy is more heavily favored.

- Chemists are very silly people.

13.3 Temperature Scales

As we will see in increasing detail as we get further into the chapter, heat is just another form of energy, and temperature, as a measure of how hot a body is, should therefore properly be expressed in energy units. In a more sensible world the weather report would, for example, forecast a high of a comfortable $\frac{1}{40}$ eV, and in the context of cosmology it is in fact customary to specify temperatures in the early universe in units of GeV (1 GeV $= 10^9$ eV).[9] Historically, however, people began measuring heat before they had any idea what it was, and we are therefore, in much the same way that we are stuck with the qwerty keyboard, stuck with a bunch of silly "temperature" scales.

The Fahrenheit scale most familiar in the United States is also the least sensible: on the Fahrenheit scale, water freezes at 32°F and boils at 212°F. This scale was proposed by Daniel Gabriel Fahrenheit in 1724, apparently without irony while he was stone cold sober. To this day the logic behind this scale remains one of the great mysteries in science.

Anders Celsius took a better stab at it in 1742 with his Celsius scale, which is set up so that water freezes at 0°C and boils at 100°C.[10] Because of this temperature difference of 100 degrees, the Celsius scale is also known as the centigrade scale. Zero and 100 are of course much nicer than 32 and 212, but Celsius could still be faulted for his placement of the scale's zero: 0°C is far from the lowest possible temperature.

[8] You may have noticed that tables of equilibrium constants do in fact state the temperature to which the listed values of K correspond.

[9] Recall that eV stands for electron volts and that 1 eV $= 1.60217653\times10^{-19}$ J. As noted in §13.1, the Boltzmann constant $k = 1.3806505\times10^{-23}$ J/K serves as the conversion factor between degrees and energy units, with the energy given by kT.

[10] The modern definition of the Celsius scale is more complicated, but the way Celsius thought of it is good enough for us.

A bloke named Lord Kelvin tried to improve the situation still further in 1848 by proposing the Kelvin scale, which uses degrees of the same size as the centigrade degree, so that there are still 100° between the freezing and boiling points of water, but puts the zero of the scale at the coldest possible temperature. This temperature, at which all molecular motion ceases, is called absolute zero, and temperature scales that have their zeros at absolute zero are called absolute temperature scales. As it turns out, absolute zero is about $-273.15°\text{C} = -459.67°\text{F}$.

As we have above, it is conventional to denote temperatures in degrees Fahrenheit and centigrade by °F and °C, and for many years the notation °K was similarly used for degrees Kelvin. It seems, however, that some influential people were offended by the eminent sensibleness of this usage, with the result that the rest of us have for some decades now been forced to denote temperatures in degrees Kelvin simply with a K. Those of you who wrote °K may even have been beaten by your chemistry teachers. It is as though they want to apotheosize the Kelvin scale as a fundamental measure of temperature. Yet another example of what silly people chemists are. There is nothing fundamental about the Kelvin scale; as a means of measuring temperature — which, as we've already noted, should properly be measured in energy units — , it's almost as silly as the Fahrenheit and centigrade scales. But you have to pick your battles carefully, and rather than fight city hall over this one, we'll just go along with the prevailing usage and denote temperatures in degrees Kelvin as though they were characters in a Kafka novel.

Anyway, as you should be able to glean from their definitions above, the relations between the temperature T_F, T_C, and T_K in degrees Fahrenheit, Celsius, and Kelvin, respectively, are

$$T_C = \tfrac{5}{9}(T_F - 32) \qquad\qquad T_K = T_C + 273.15 \tag{13.6}$$

And that's about all that need be said about that.[11]

13.4 Heat Energy & Changes of Temperature & Phase

As a body is heated, its temperature rises. Not an observation likely to win a Nobel Prize, but there you have it: the body gets hotter as the added

[11] Except perhaps to note that there are several other temperature scales in addition to Fahrenheit, Celsius, and Kelvin. The Rankine scale, proposed in 1859 by William John Macquorn Rankine, is an absolute scale with the same size of degree as the Fahrenheit scale. But in spite of the lordly gravity of its propounder's name, the Rankine scale has never seen much use.

heat energy increases the thermal motion of the molecules within the body. An exception to this is a body undergoing a phase transition, that is, a transition from one of its solid, liquid, and gaseous states to another: during these transitions, heat added to or taken from the body contributes only to a change in the potential energy associated with the forces that hold the molecules in liquid or solid form, and there is no change in temperature until the phase change is complete. When ice is heated, for example, its temperature rises until it reaches the melting point at 0°C. At that point all additional heat goes into loosening the bonds that hold the water molecules rigidly in solid form. Only when all the bonds have been loosened and the entire sample has melted into liquid water will continued heating again raise the temperature.

In quantitative relations, we will use Q to denote the heat energy *absorbed* by a body; if a body gives off heat, Q has a negative value. It should seem plausible that in between phase transitions changes in temperature are given by

$$Q = mc\Delta T \tag{13.7}$$

where m is the mass of the sample and c, known as the specific heat, depends on the substance of which the sample is composed. Eq. (13.7) is simply stating that the temperature rise of a given sample is proportional to the heat energy added (or, in the case of negative Q, that the temperature drop is proportional to the heat removed). In this proportionality, the factor of the mass m takes into account how much of the substance we have: if we have twice as much of a given substance, it should take twice as much heat energy to produce the same temperature increase. The remaining constant of proportionality, the specific heat c, takes into account that the addition of the same amount of heat energy to samples of the same mass may — and in fact almost always does — result in different temperature changes for different substances. The specific heat may also differ for different phases of the same substance — for example, the specific heats of ice, water, and steam differ. In fact, even within a phase the specific heat is a function of temperature,[12] so that eq. (13.7) strictly applies only to infinitesimal changes and would more precisely be written as

$$dQ = mc\,dT$$

To determine the change in temperature due to a finite added heat energy Q we would have to know c as a function of temperature and solve

$$Q = m\int_{T_0}^{T} c\,dT$$

[12]Not to mention pressure, etc. Life is complicated.

for the final temperature T. The temperature dependence of the specific heat is, however, not infrequently weak enough that it can be neglected to a good approximation, and we will for simplicity always assume that the specific heat for a given phase of a substance is constant.

For phase changes, it should likewise seem plausible that

$$Q = \pm mL, \tag{13.8}$$

where m is once again the mass of the sample and L, known as the latent heat, depends on the substance of which the sample is composed. Eq. (13.8) is simply stating that the heat energy required for a phase change should be proportional to the mass of the sample: if we have twice as much of a given substance, it should take twice as much heat energy to bring about the phase change. Unlike eq. (13.7), however, eq. (13.8) does not automatically take into account the direction of the change: to melt a solid or vaporize a liquid, the substance must absorb heat, so that Q is positive and we want $Q = +mL$; to solidify a liquid or condense a gas, heat must be removed from the substance, so that Q is negative and we want $Q = -mL$. Since we will not be worried about the transition directly from solid to gas that occurs in sublimation, the latent heat L will come for us in just two flavors: latent heat of fusion for transitions between solid and liquid (melting or solidification), and the latent heat of vaporization for transitions between liquid and gas (vaporization or condensation). The value of the latent heat is determined by the potential energy corresponding to the forces that hold the substance's molecules together in liquid or solid form and therefore differs from one substance to another.

The heat energy Q is measured in the same units as any other energy. For us, that will often mean Joules. Historically, however, there is another common unit for measuring heat energies: the calorie (cal). Basically, 1 cal of heat energy will raise the temperature of 1 g of water by 1°C. More precisely, the calorie is defined as the heat energy necessary to raise the temperature of 1 g of water at 1 atm of pressure from 3.5°C to 4.5°C.[13] In terms of Joules, 1 cal $\approx$ 4.186 J.[14] The calorie is a convenient unit because water is so common and so often constitutes the dominant thermal component of solutions and substances.

In the context of food, the word *calorie* is conventionally used for *kilocalorie*: when you see "200 calories" on a food label, it should really read "200 kilocalories." Presumably this abuse started when people who wrote

[13] Actually, this is just one of several inconsistent definitions of a calorie — inconsistent in the sense that different definitions yield slightly different values for the calorie. The above definition is centered at 4°C because that is the temperature at which water is densest and at which the specific heats at constant pressure and constant volume are therefore equal.

[14] We give this conversion as an approximation because, as noted in the preceding footnote, the precise value of the calorie depends on which of its several definitions is used.

diet books, confronted with what apparently was to them the disconcertingly strange and unfamiliar prefix *kilo*, chose simply to drop the prefix and hope that by ignoring it they could make it go away. Then, as happens so often in the course of human events, if enough people commit a wrong, it comes to be considered a right. Perhaps in your lifetime, people will even be burned at the stake for using the term *kilocalorie*.[15]

As a rather artificial application of this whole business of specific and latent heats, consider a system thermally isolated — by means of thermal insulation or some other mechanism — so that it cannot exchange any heat with its surroundings. For such a system, only heat exchanges among the various parts of the system are possible, and no net heat is absorbed or given off by the system as a whole. Thus

$$\sum_{\text{parts}} Q_{part} = 0 \tag{13.9}$$

for a thermally isolated system. When bodies at differing temperatures are combined in such an isolated environment, the Q's for the exchanges of heat will consist of temperature changes of the form $mc\Delta T$ and phase changes of the form $\pm mL$, and we can then use eq. (13.9) to determine the equilibrium temperature and state to which the system settles down as the temperatures of its various parts equalize.

Suppose, for example, that you are splashing around in a tub full of Mr. Bubble when a 200 kg, 800°C meteor crashes through the roof and lands in the water. If for simplicity we take you (who are mostly water) and the tub water together to be equivalent to 300 kg of 40°C water, then we have

[15] Actually, times have changed. In today's "civilized" society, instead of putting them through a few minutes of agony at the stake, we shun and ostracize those regarded as politically incorrect, with the object of making their lives miserable until they concede that $2 + 2 = 5$ or 3 or whatever the current Party line is. While not as spectacularly entertaining as a good auto-da-fé or feeding someone to the lions in the Colosseum, it serves the same purpose, and has the additional advantage of allowing the mob to feel morally superior. Usually these exercises are conducted in the name of "tolerance" — which, by a most marvelous sort of Orwellian double-speak, has come to stand for the most extreme sort of intolerance. "If you want a picture of the future, imagine a boot stamping on a human face — forever."

And while we're on the subject of double-speak, note how political correcticians refer to themselves as liberal when in fact, quite the opposite, they are extremely intolerant of any views that differ from their own: in a new incarnation of McCarthyism, they ostracize and persecute those who think differently from them and seek, in their arrogance and hypocrisy, to control not only people's actions and words, but even their thoughts. For a true liberal, freedom of thought and expression is a fundamental and guiding principle; the true liberal will challenge your point of view but respect both it and your right to express it. Sadly, such liberals have always been few and far between. And yet there has never been a greater need of them than now, when political correctness has become such a grave threat to all that raises human existence above the level of rats and roaches.

a two-body system consisting initially of 300 kg of 40°C water and 200 kg of 800°C solid iron meteor, so that [16]

$$Q_{water} + Q_{meteor} = 0 \tag{13.10}$$

Since the iron meteor, which is already solid, will simply cool off without undergoing any phase change, its Q is of the form $mc\Delta T$. Dealing with the water is a bit less certain: we know that it will heat up, but will it remain liquid or will it partly or completely boil off? If we make the tentative assumption that some but not all of the water boils off, then the Q for the water consists of an $mc\Delta T$ to raise all 300 kg of it to the 100°C boiling point and an mL for however much of the water boils off. Under this assumption, we also know that the final state will be a mixture of liquid water and steam, which must therefore be at 100°C. If we denote by m_{vap} the mass of water that boils off and note that

$$\begin{aligned} c &= 1.0\,\text{cal/g°C} \quad && \text{(liquid water)} \\ c &= 0.16\,\text{cal/g°C} && \text{(solid iron)} \\ L &= 540\,\text{cal/g} && \text{(vaporization of water)} \end{aligned}$$

then eq. (13.10) expands into

$$\begin{aligned} 0 &= Q_{water} + Q_{meteor} \\ &= \Big[[mc\Delta T]_{40^\circ \text{ to } 100^\circ} + m_{vap}L\Big]_{water} + \Big[[mc\Delta T]_{800^\circ \text{ to } 100^\circ}\Big]_{iron} \\ &= 300(1.0)(100-40) + m_{vap}(540) + 200(0.16)(100-800) \end{aligned}$$

which yields

$$m_{vap} = 8.1\,\text{kg}$$

So, under the assumption that some of the water boils off, we find that 8.1 kg of the water is turned to steam.

That our result for the mass of water that boils off is sensible validates the assumption we used to set up the calculation. Thermal equilibrium calculations are frequently like this: we have to make some assumption about what happens in order to set up the relation among the Q's.[17] If the assump-

[16] Of course, treating this system as thermally isolated is a terrible approximation: it is in direct thermal contact both with the tub and the surrounding air, to which heat will fairly rapidly be lost both by direct transfer and through evaporation — as those of you who have tried to take a long hot bath, with or without the Mr. Bubble and the meteor, know from experience.

[17] Alternatively, we could work out the Q contributions from, as it were, the outside in, by looking first at the Q's for cooling the hottest bodies down, and heating the coldest bodies up, to the temperatures at which they undergo phase changes, then working out the Q's for the phase changes, then working out more temperature-change Q's, and so on, until the two ends meet. This method has the advantage that you never make a bad assumption about the final state but the disadvantage that the calculation is more involved.

tion is correct, we will get a sensible result; if not, the result we get will not make sense but will hint at the correct assumption. If, for example, we had instead assumed that the water heats up but all remains liquid, we would have had only an $mc\Delta T$ term for the water, and our unknown would be the final equilibrium temperature T_{eq}:

$$\begin{aligned} 0 &= Q_{water} + Q_{meteor} \\ &= \Big[[mc\Delta T]_{40^\circ \text{ to } T_{eq}}\Big]_{water} + \Big[[mc\Delta T]_{800^\circ \text{ to } T_{eq}}\Big]_{iron} \\ &= 300(1.0)(T_{eq} - 40) + 200(0.16)(T_{eq} - 800) \end{aligned}$$

This would yield

$$T_{eq} = 113^\circ\text{C}$$

which of course doesn't make sense: we can't have liquid water at over 100°C. But getting an equilibrium temperature above 100°C tells us that there is more than enough heat to bring the system to 100°C and that therefore at least some of the water boils off.

Before moving on we should note that perfect thermal isolation is actually impossible to achieve: the molecular motion within all bodies that are above absolute zero causes them to emit electromagnetic radiation known as blackbody radiation. While invisible at room temperature, this is the radiation that causes fireplace pokers and other things to glow red or even white when they are really hot — hence the expressions *red hot* and *white hot*. The spectrum of this blackbody radiation depends on temperature: while bodies at the same temperature share the same blackbody radiation spectrum, so that each reabsorbs as much blackbody radiation energy from the others as it gives off, bodies at different temperatures have different spectra, with the result that the hotter bodies shed more blackbody radiation energy than they absorb back from the colder bodies. Even a total vacuum between a system and its surroundings cannot prevent heat transfer due to blackbody radiation. Consequently the only system that is perfectly thermally insulated is the universe itself: there is nothing outside of it with which it can exchange heat. But good thermal insulation can limit heat exchange with the surroundings enough that a less grand system can, at least for reasonably short periods of time, be considered thermally isolated to a very good approximation.

13.5 Ideal Gases

The molecules of an ideal gas are point-like and undergo only simple billiard-ball–like elastic collisions with each other and with the walls of any vessel

confining them. Perhaps the most salient characteristic of such gases is that they don't exist: the behavior of real gases, even noble gases, is not really this simple. But if there were such a thing as ideal gases, they would obey what, by a fortuitous coincidence, is known as the ideal-gas law. In a previous life in chemistry, you probably expressed the ideal-gas law as

$$pV = nRT \tag{13.11}$$

where p, V, and T are the pressure, volume, and temperature of the gas, respectively; n is the amount of gas, measured in a unit known as a mole that makes sense only to chemists; and $R = 8.3144727\ \mathrm{J/mol{\cdot}K}$ is a constant of proportionality known as the gas constant. A much more sensible way to express the ideal-gas law is

$$pV = NkT \tag{13.12}$$

where N is the number of molecules and $k = 1.3806505 \times 10^{-23}\ \mathrm{J/K}$ is the Boltzmann constant. Since, as noted in §13.1, the Boltzmann constant is just a sort of conversion factor between the artificial units of degrees in which temperature is conventionally measured and the energy units in which it should properly be measured, it would be even more sensible to use an energy in place of kT, but we live in an imperfect world.

Anyway, rather than waste time crying about the world's imperfections, let us get on with our lives. As you can see by comparing the right-hand sides of eqq. (13.11) and (13.12), the relation between the Boltzmann constant k and the gas constant R is

$$nR = Nk$$

Since the number n of moles is the number N of molecules divided by Avogadro's number N_0, we can rewrite this as

$$\frac{N}{N_0} R = Nk$$

which yields

$$R = N_0 k$$

The gas constant, Avogadro's number, and the Boltzmann constant are therefore not independent; if we know the values of any two, we can determine the value of the third.

The ideal-gas law is an example of an equation of state that governs the state of a system: it tells us that the state of a given sample of N molecules of an ideal gas is determined by the values of the pressure p, temperature T, and volume V. The variables p, T, and V are in turn known as state variables: the value of each of these variables depends only on the state of the sample of gas and not on the history of the gas, that is, not on how the gas was brought into that state. These variables uniquely specify the macroscopic state of the

system even though there are an astronomical number of microscopic states — sets of particular velocities and locations for each of the molecules in the sample of gas — that could each give rise to that same macroscopic state. State variables can be either independent or dependent, and in this case the gas law is telling us that only two of p, V, and T are independent; the third is then determined by $pV = NkT$.

We will now derive the ideal-gas law by working out the Newtonian kinematics of the billiard-ball–like molecules of an ideal gas.[18] Suppose we have N molecules of an ideal gas at pressure p and temperature T confined to a volume V by the walls of a vessel. The first step in relating kinematic quantities like force, momentum, mass, and velocity to pressure and volume is to apply the definition of pressure,[19]

$$\text{pressure} = \frac{\text{force}}{\text{area}}$$

to an infinitesimal patch of area dA on the wall of the vessel: if we use P for momentum to avoid confusion with the pressure p, then, since $F = dP/dt$, the contribution to the pressure from the collisions of the gas molecules with the wall is of the form

$$p = \frac{dF}{dA} = \frac{dP/dt}{dA} \tag{13.13}$$

Compared to the molecules striking it, the wall is effectively infinitely massive. Recall now from §6.7.1 that in a one-dimensional elastic collision with an infinitely massive stationary target, the velocity of the incident body simply reverses direction. In our present three-dimensional context, the components of the molecule's velocity parallel to the wall of the vessel will be unchanged and the component perpendicular to the wall will be reversed.[20] If we denote this perpendicular component by $v_\perp$, then as a result of impact with the wall

$$\Delta v_\perp = v_\perp - (-v_\perp) = 2v_\perp$$

If the mass of each molecule is m and we denote by $dN_{v_\perp}$ the number of molecules moving with velocity $v_\perp$ that strike the wall during the time interval dt, the change in the momentum of these molecules is their total mass $dN_{v_\perp}\, m$ times their change in velocity $2v_\perp$:

$$dP = dN_{v_\perp}\, m\, (2v_\perp) = 2mv_\perp\, dN_{v_\perp}$$

[18] Again, for those of you who like jargon, these sorts of kinematical calculations constitute what is known as kinetic theory.

[19] The MKS unit of pressure is the Pascal (Pa): $1\,\text{Pa} = 1\,\text{N/m}^2$. A common alternative unit is the atmosphere (atm): $1\,\text{atm} = 1.01325 \times 10^5\,\text{Pa}$, or, in English units, $14.70\,\text{lb/in}^2$. Units of millimeters of mercury (mmHg) are discussed on p.622, where pressure is introduced in the context of fluid dynamics, but we won't be needing those units for thermal physics.

[20] The three-dimensional case was actually worked out in problem #29 on p.334.

If we use this in eq. (13.13), the contribution $dp_{v_\perp}$ of these $dN_{v_\perp}$ molecules to the pressure p is

$$dp_{v_\perp} = \frac{2mv_\perp \, dN_{v_\perp}}{dA \, dt} \tag{13.14}$$

The next step is to relate the right-hand side of eq. (13.14) to volume by considering the infinitesimal volume $dV = dA \, d\ell$ of perpendicular depth $d\ell$ adjacent to the patch dA of the wall. Of the total number $dN_{v_\perp, dV}$ of molecules moving with perpendicular velocity $v_\perp$ within the volume dV, only half will be moving toward the wall, and of that half only those within a perpendicular distance $d\ell_\perp = v_\perp \, dt$ of the wall will reach it during the time interval dt. Stated another way, of the half of these molecules headed toward the wall, the fraction within the volume dV that will reach the wall during the time interval dt is the ratio of the distance $d\ell_\perp = v_\perp \, dt$ that they travel toward the wall to the depth $d\ell$ of dV. Thus

$$dN_{v_\perp} = dN_{v_\perp, dV} \left(\frac{1}{2}\right) \left(\frac{v_\perp \, dt}{d\ell}\right)$$

Using this in eq. (13.14) yields

$$\begin{aligned} dp_{v_\perp} &= \frac{2mv_\perp}{dA \, dt} \, dN_{v_\perp, dV} \left(\frac{1}{2}\right) \left(\frac{v_\perp \, dt}{d\ell}\right) \\ &= \frac{m \, dN_{v_\perp, dV} \, v_\perp^2}{dA \, d\ell} \\ &= \frac{m \, dN_{v_\perp, dV} \, v_\perp^2}{dV} \end{aligned}$$

This can be made a bit more digestible by rewriting it in terms of the total number of molecules dN within the volume dV and the fraction

$$f_{v_\perp} = \frac{dN_{v_\perp, dV}}{dN}$$

of these molecules that are moving with perpendicular velocity $v_\perp$: substituting $f_{v_\perp} dN$ for $dN_{v_\perp, dV}$, we have

$$dp_{v_\perp} = \frac{m f_{v_\perp} dN \, v_\perp^2}{dV} \tag{13.15}$$

While this might not seem like much of an improvement, note that dN/dV, the number of molecules per unit volume, is uniform throughout the gas and can therefore be replaced by N/V. And when we integrate $dp_{v_\perp}$ over all possible values of $v_\perp$ in order to get the total pressure p, the integration of

$$f_{v_\perp} v_\perp^2$$

is just a weighted average that will yield the average value of $v_\perp^2$. If we denote this average value by $\overline{v_\perp^2}$, eq. (13.15) thus yields

$$p = \frac{mN\overline{v_\perp^2}}{V}$$

and hence

$$pV = Nm\overline{v_\perp^2} \tag{13.16}$$

As we will now show, $m\overline{v_\perp^2} = kT$, so that we do in fact have $pV = NkT$.

To obtain a result for the average kinetic energy of the gas molecules and hence for $m\overline{v_\perp^2}$, we turn to statistical mechanics. Recall that the basic tenet of statistical mechanics is that the probability of a system being in a state i of energy E_i is proportional to $e^{-\beta E_i}$, where $\beta = 1/kT$. If we denote the constant of proportionality by $\mathcal{N}$, then the absolute probability of being in state i is

$$\mathcal{N}e^{-\beta E_i}$$

Since the system is certainly in some state or other, the sum of these probabilities over all states i must be unity:

$$\sum_i \mathcal{N}e^{-\beta E_i} = 1$$

Since $\mathcal{N}$ is a constant, we can pull it outside of the sum to obtain

$$\mathcal{N}\sum_i e^{-\beta E_i} = 1$$

and hence

$$\mathcal{N} = \frac{1}{\sum\limits_i e^{-\beta E_i}} \tag{13.17}$$

This animal $\mathcal{N}$ is known as the normalization, and the absolute probability of being in state i is thus[21]

$$\frac{e^{-\beta E_i}}{\sum\limits_j e^{-\beta E_j}} \tag{13.18}$$

The average value $\overline{X}$ of a quantity X will be the average of its values X_i over the possible states i of the system, each weighted by the probability (13.18) of being in state i:

$$\overline{X} = \sum_i X_i \frac{e^{-\beta E_i}}{\sum\limits_j e^{-\beta E_j}}$$

[21] We have changed the dummy index in the sum for $\mathcal{N}$ from i to j in order to avoid confusion with the index i already used in the $e^{-\beta E_i}$.

In particular, the average energy of the system will be

$$\overline{E} = \sum_i E_i \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}} \tag{13.19}$$

The partition function Z of a system is defined to be

$$Z = \sum_i e^{-\beta E_i} \tag{13.20}$$

where the E_i are the energies of the various possible states of the system. The partition function is so called because it tells us how the probabilities of these states are partitioned, in the sense of divvied up among them. It has the handy property that in terms of it the average energy of the system is given by

$$\overline{E} = -\frac{\partial}{\partial\beta} \ln Z \tag{13.21}$$

To see that this is so, we need merely expand the right-hand side by the chain rule:

$$\begin{aligned} -\frac{\partial}{\partial\beta} \ln Z &= -\frac{\partial}{\partial\beta} \ln \sum_i e^{-\beta E_i} \\ &= -\frac{1}{\sum_j e^{-\beta E_j}} \frac{\partial}{\partial\beta} \sum_i e^{-\beta E_i} \\ &= -\frac{1}{\sum_j e^{-\beta E_j}} \sum_i \left(-E_i\, e^{-\beta E_i}\right) \\ &= \sum_i E_i \frac{e^{-\beta E_i}}{\sum_j e^{-\beta E_j}} \end{aligned}$$

which, by eq. (13.19), is indeed $\overline{E}$.

The energy of an ideal gas in general consists of translational, rotational, and vibrational contributions — the latter corresponding to oscillations of the atoms in a molecule relative to each other just like those you would get if the molecule were made up of little billiard balls connected by springs. If for simplicity we assume that our gas is monatomic, so that the molecules are point-like, then we don't have to worry about rotational and vibrational modes, and the energy of each molecule is due solely to its translational kinetic energy

$$E = \tfrac{1}{2}mv^2 = \tfrac{1}{2}m(v_x^2 + v_y^2 + v_z^2)$$

In terms of the partition function, the average energy of each molecule of a monatomic ideal gas is consequently obtained by integrating over all the

possible values of the components v_x, v_y, and v_z of the molecule's velocity:[22]

$$\begin{aligned}
\overline{E} &= -\frac{\partial}{\partial\beta}\ln Z \\
&= -\frac{\partial}{\partial\beta}\ln\int_{-\infty}^{\infty}dv_x\int_{-\infty}^{\infty}dv_y\int_{-\infty}^{\infty}dv_z\,e^{-\beta\cdot\frac{1}{2}mv^2} \\
&= -\frac{\partial}{\partial\beta}\ln\int_{-\infty}^{\infty}dv_x\int_{-\infty}^{\infty}dv_y\int_{-\infty}^{\infty}dv_z\,e^{-\beta\cdot\frac{1}{2}m(v_x^2+v_y^2+v_z^2)} \\
&= -\frac{\partial}{\partial\beta}\ln\Bigl(\int_{-\infty}^{\infty}dv_x\,e^{-\beta\cdot\frac{1}{2}mv_x^2}\int_{-\infty}^{\infty}dv_y\,e^{-\beta\cdot\frac{1}{2}mv_y^2} \\
&\qquad\qquad \int_{-\infty}^{\infty}dv_z\,e^{-\beta\cdot\frac{1}{2}mv_z^2}\Bigr)
\end{aligned} \tag{13.22}$$

To carry out each of the three identical integrals inside the logarithm, we first make changes of variables like

$$u = \sqrt{\frac{\beta m}{2}}\,v_x$$

to get rid of the extraneous factors of β and m in the exponential:

$$\begin{aligned}
\int_{-\infty}^{\infty}dv_x\,e^{-\beta\cdot\frac{1}{2}mv_x^2} &= \int_{-\infty}^{\infty}dv_x\,\exp\left(-\left(\sqrt{\frac{\beta m}{2}}\,v_x\right)^2\right) \\
&= \sqrt{\frac{2}{\beta m}}\int_{-\infty}^{\infty}d\left(\sqrt{\frac{\beta m}{2}}\,v_x\right)\exp\left(-\left(\sqrt{\frac{\beta m}{2}}v_x\right)^2\right) \\
&= \sqrt{\frac{2}{\beta m}}\int_{-\infty}^{\infty}du\,e^{-u^2}
\end{aligned} \tag{13.23}$$

The famous mathematical trick to carrying out this commonly occurring remaining integral is to note that its square can be converted into an integration over a plane that can be carried out in polar coordinates:

$$\begin{aligned}
\int_{-\infty}^{\infty}dx\,e^{-x^2}\int_{-\infty}^{\infty}dy\,e^{-y^2} &= \int\limits_{xy\,\text{plane}}dx\,dy\,e^{-(x^2+y^2)} \\
&= \int\limits_{xy\,\text{plane}}dA\,e^{-(x^2+y^2)}
\end{aligned}$$

[22] In general discussions like that above, the partition function is conventionally written as a sum over discrete energy levels E_i. The Newtonian mechanical energies with which we are dealing are, however, functions of continuous parameters — the velocity v in $\frac{1}{2}mv^2$ or the like — , so that instead of a discrete sum we have a continuous integral.

$$
\begin{aligned}
&= \int\limits_{xy\,\text{plane}} r\,dr\,d\theta\,e^{-r^2} \\
&= \int_0^{2\pi} d\theta \int_0^{\infty} r\,dr\,e^{-r^2} \\
&= 2\pi \int_0^{\infty} \tfrac{1}{2}\,d(r^2)\,e^{-r^2} \\
&= 2\pi \cdot \tfrac{1}{2}\Big[-e^{-r^2} \Big]_0^{\infty} \\
&= 2\pi \cdot \tfrac{1}{2}\Big(0 - (-1)\Big) \\
&= \pi
\end{aligned}
$$

Thus

$$\int_{-\infty}^{\infty} du\,e^{-u^2} = \sqrt{\pi}$$

and hence eq. (13.23) works out to

$$\int_{-\infty}^{\infty} dv_x\,e^{-\beta\cdot\frac{1}{2}mv_x^2} = \sqrt{\frac{2\pi}{\beta m}} \tag{13.24}$$

and hence hence eq. (13.22) to

$$
\begin{aligned}
\overline{E} &= -\frac{\partial}{\partial\beta} \ln\left(\sqrt{\frac{2\pi}{\beta m}}\right)^3 \\
&= -\frac{\partial}{\partial\beta} \ln\left(\frac{\beta m}{2\pi}\right)^{-\frac{3}{2}} \\
&= \frac{3}{2}\frac{\partial}{\partial\beta}\left(\ln\frac{\beta m}{2\pi}\right) \\
&= \frac{3}{2}\frac{\partial}{\partial\beta}\left(\ln\beta + \ln\frac{m}{2\pi}\right) \\
&= \frac{3}{2}\frac{\partial}{\partial\beta}\ln\beta + 0 \\
&= \frac{3}{2}\frac{1}{\beta}
\end{aligned}
$$

or, since $\beta = 1/kT$, to

$$\overline{E} = \tfrac{3}{2}kT \tag{13.25}$$

The total energy of N molecules is therefore $\frac{3}{2}NkT$, modulo statistical fluctuations on the order of $1/\sqrt{N}$ that will be negligibly small for macroscopic

samples of gas. This total energy of the sample is called its internal energy and is conventionally denoted by U:

$$U = \tfrac{3}{2}NkT \tag{13.26}$$

Eq. (13.26) is a particular example of the general result that when the dependence of the energy on a variable is quadratic, as it is in $\frac{1}{2}mv^2$, there is $\frac{1}{2}kT$ of internal energy per degree of freedom, with a degree of freedom being essentially a way — in an abstract sense, a dimension — along which the system can move and thus "suck up" energy: the molecules of a monatomic ideal gas are free to move along any of the three spatial directions, and consequently each molecule has an average energy of $3 \cdot \frac{1}{2}kT$. We will offer one further example of this general result at the end of this section.

First, however, we need to tidy up some unfinished business: in our derivation of the ideal-gas law, we said that $m\overline{v_\perp^2}$ was equal to kT. As the component of the velocity along a particular spatial axis — the axis perpendicular to the patch dA of the wall — , $v_\perp$ is no different from v_x, v_y, v_z, or the component of v along any other axis: this direction of motion constitutes a degree of freedom, and the average value of $\frac{1}{2}mv_\perp^2$ will therefore be $\frac{1}{2}kT$. Thus

$$m\overline{v_\perp^2} = kT$$

as claimed.[23]

We now end with the one-dimensional harmonic oscillator as a further illustration of the general result that there is $\frac{1}{2}kT$ of energy per degree of freedom. The energy of a mass m moving along an x axis under the influence of a harmonic potential $\frac{1}{2}kx^2$ is

$$E = \tfrac{1}{2}mv^2 + \tfrac{1}{2}kx^2$$

There are thus two degrees of freedom, x and v, so that we expect

$$\overline{E} = 2 \cdot \tfrac{1}{2}kT = kT$$

The partition function now involves an integration over all the possible values of both the velocity v and location x of the mass m, so that by eqq. (13.20) and (13.21) the average energy is

$$\begin{aligned}\overline{E} &= -\frac{\partial}{\partial\beta}\ln\int_{-\infty}^{\infty}dv\int_{-\infty}^{\infty}dx\,e^{-\beta(\frac{1}{2}mv^2+\frac{1}{2}kx^2)} \\ &= -\frac{\partial}{\partial\beta}\ln\left(\int_{-\infty}^{\infty}dv\,e^{-\beta\cdot\frac{1}{2}mv^2}\int_{-\infty}^{\infty}dx\,e^{-\beta\cdot\frac{1}{2}kx^2}\right)\end{aligned}$$

[23] If you aren't convinced of this, you can work it out from scratch by repeating our calculation of $\overline{E}$ in eq. (13.22) for $\frac{1}{2}mv_\perp^2$ by itself.

Each of these integrals is identical to that of eq. (13.23), except that in the integral over x we have k instead of m. From eq. (13.24) we see that these integrals therefore yield

$$\begin{aligned}
\overline{E} &= -\frac{\partial}{\partial\beta}\ln\left(\sqrt{\frac{2\pi}{\beta m}}\,\sqrt{\frac{2\pi}{\beta k}}\right) \\
&= -\frac{\partial}{\partial\beta}\ln\left(\frac{1}{\beta}\frac{2\pi}{\sqrt{km}}\right) \\
&= -\frac{\partial}{\partial\beta}\left(\ln\frac{1}{\beta} + \ln\frac{2\pi}{\sqrt{km}}\right) \\
&= -\frac{\partial}{\partial\beta}\left(-\ln\beta + \ln\frac{2\pi}{\sqrt{km}}\right) \\
&= \frac{1}{\beta} + 0 \\
&= kT
\end{aligned}$$

Word.

13.6 Processes, Cycles, & the First Law

The first law of thermodynamics is just conservation of energy phrased in thermodynamic terms. In words, it states that the change in the internal energy of a system equals the heat absorbed by the system less the work done by the system. The reasoning behind this is just common sense: if we put heat energy into a system, some of that may reappear in the form of work done by the system, and the rest, simply because there is no other way to account for it, must still be stored internally in the system in the form of molecular kinetic and potential energy. With the notation

$$\begin{aligned}
U &= \text{internal energy of the system} \\
Q &= \text{heat energy } \textit{absorbed} \text{ by the system} \\
W &= \text{work done } \textit{by} \text{ the system}
\end{aligned}$$

the first law takes the form

$$dU = dQ - dW \tag{13.27a}$$

for infinitesimal energy transfers and

$$\Delta U = Q - W \tag{13.27b}$$

for finite energy transfers. Note that dQ (or Q) is negative if the system is *giving off* heat and that dW (or W) is negative if work is being done by you or some other agent *on* the system.

A process is any operation or action that changes the state of a system. For simplicity we will restrict ourselves to systems consisting of samples of ideal gases, but that restriction won't prove severe — it turns out that you can do an awful lot with ideal gases. At any rate, when a sample of an ideal monatomic gas is subjected to a continuous process that takes it from an initial state A to a final state B, it passes through an infinite succession of intermediate states, each differing infinitesimally from the preceding state. Each of the initial, intermediate, and final states[24] has a definite value of each of the state variables p, T, and V and obeys the equation of state

$$pV = NkT$$

along with

$$U = \tfrac{3}{2}NkT$$

In addition, the transitions between successive states all obey the first law,

$$dU = dQ - dW$$

For our ideal monatomic gas, you can see from $U = \frac{3}{2}NkT$ that the value of U depends only on the temperature T and is therefore uniquely determined by the state of the gas. This turns out to be true quite generally: the internal energy U of any system is a state variable, the value of which depends only on the state of the system and not on how the system was brought into that state. The overall change ΔU_{AB} in internal energy between an initial state A and a final state B therefore depends only on the states A and B and not on the process by which the system was brought from A to B, so that for any process that takes the system from state A to state B we have

$$\Delta U_{AB} = U_B - U_A$$

We make a big deal out of this because the same is *not* true of Q and W: as we will see explicitly below, the heat energy absorbed and the work done by the system depend not only on the initial and final states of the system, but on the process by which the system was brought from the initial to the final state. Q and W are *not* state variables.[25] This will be most apparent

[24] Actually, we're being a bit sketchy here in the interest of simplicity: throughout the chapter we make the tacit assumption that these states are also equilibrium states; for nonequilibrium states variables like temperature and pressure may not even be well defined. We will address this issue explicitly at the end of §13.6.4.

[25] For this reason, some authors distinguish differentials of state and non-state variables by, for example, putting a bar through the d for the latter, so that eq. (13.27a) would be written $dU = đQ - đW$. We won't use this notation simply because it can prove rather cumbersome.

when we deal with cycles, that is, with processes that return the system to its original state: Since the final state is the same as the initial state, the overall change in the internal energy must vanish around a cycle, so that

$$\Delta U_{cycle} = 0$$

or, phrased in terms of the infinitesimal changes around a continuous cycle,

$$\oint_{\text{cycle}} dU = 0$$

This is not, however, true of Q and W: we will see that even though it returns to its original state, the system may have absorbed or shed a nonzero net amount of heat energy and may have done (or have had done on it) a nonzero net amount of work.

Before we get into examples to illustrate this, however, it will be helpful to make a few general observations about the ideal gases with which we will be working. First, as we noted earlier, only two of the variables p, T, and V are independent; the third is then determined by the equation of state $pV = NkT$. The state of a sample of gas can therefore be uniquely specified by the values of just p and V, and we can think of each state as a point (p, V) in the pV plane.[26] A continuous process that takes a sample of gas from an initial state A through a succession of intermediate states to a final state B can then be represented graphically by a continuous curve in the pV plane from point (p_A, V_A) to point (p_B, V_B). A cycle that returns a sample of gas to its original state will form a closed loop.

Changes in the internal energy of the sample of gas are easily calculated using $U = \frac{3}{2}NkT$. Often the heat energy Q absorbed by the gas is most easily obtained by means of the first law, $\Delta U = Q - W$. This of course will entail our having first determined the work W done by the gas, and the key to doing so is to note that the gas does work when it expands and pushes outward against whatever surrounds it (or has work done on it when its surroundings push on and compress it). To quantify this work, consider a gas enclosed in a cylinder of cross-sectional area A fitted with a movable piston.[27] As the gas expands and pushes the piston outward a distance $d\ell$, the work dW done by the force F associated with the pressure is

$$dW = F\, d\ell$$

Using the definition $p = F/A$ of pressure in the form $F = pA$, we can re-express this as

$$dW = pA\, d\ell = p\, dV \tag{13.28}$$

[26] Since we will be plotting p on the vertical axis and V on the horizontal axis, writing points in the pV plane as (p, V) is like writing points in the xy plane as (y, x). But writing (V, p) would somehow just feel weird.

[27] Not that there would be much point in an immovable piston, but you get the idea.

where we have noted that $A\,d\ell$ is the corresponding infinitesimal increase dV in the volume of the gas. Although we have derived this result only for the special case of an expansion into an infinitesimal cylindrical volume dV, an expansion into any finite volume of any shape can be built up out of such infinitesimal cylindrical volumes, so our result is quite general. The work done for a finite change from volume from V_A to V_B is therefore

$$W = \int_{V_A}^{V_B} p\,dV$$

Geometrically, the work done by a sample of gas is thus the area under the curve corresponding to the process in the pV plane.

13.6.1 A Painfully Long Example

Consider, for example, the cycle shown in fig. (13.1). In the interest of thoroughness, we are going to beat this example to death, so it's going to be a long ride — you may want to pack a sandwich. And be sure you go potty before we get started, because we're not stopping.

Ready?

Okay: The vertical "leg" of the cycle of fig. (13.1) takes the sample of gas from pressure p_1 to a higher pressure p_2 while keeping the volume constant at V_1. By $pV = NkT$, the temperature therefore rises from

$$T_1 = \frac{p_1 V_1}{Nk}$$

to

$$T_2 = \frac{p_2 V_1}{Nk}$$

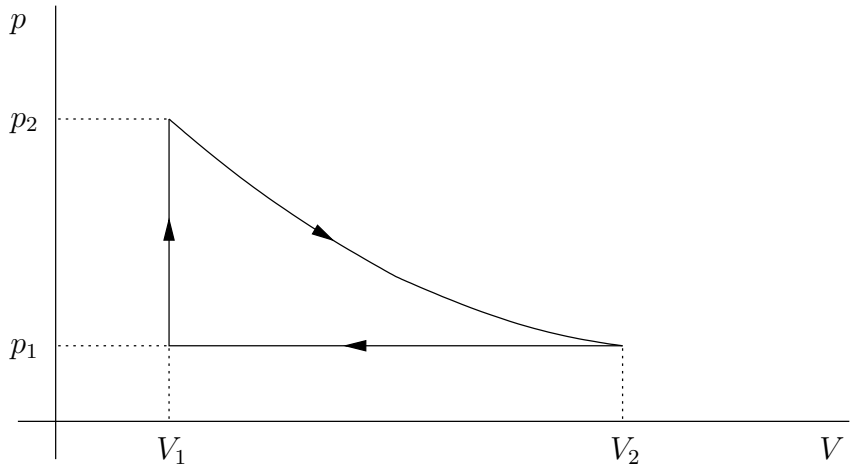


Figure 13.1: All Your Base are Belong to Us

This in turn means that the internal energy of the gas increases by

$$\Delta U = \tfrac{3}{2}Nk\Delta T = \tfrac{3}{2}Nk(T_2 - T_1) = \tfrac{3}{2}Nk\left(\frac{p_2V_1}{Nk} - \frac{p_1V_1}{Nk}\right) = \tfrac{3}{2}(p_2 - p_1)V_1$$

Since the volume is constant, $dW = p\,dV = 0$ throughout this leg and no work is done. With these results for ΔU and W, the first law gives

$$\begin{aligned} \Delta U &= Q - W \\ \tfrac{3}{2}(p_2 - p_1)V_1 &= Q - 0 \\ Q &= \tfrac{3}{2}(p_2 - p_1)V_1 \end{aligned}$$

To summarize,

$$V = V_1 \text{ (constant)} \qquad p = p_1 \to p_2 \qquad T = \frac{p_1V_1}{Nk} \to \frac{p_2V_1}{Nk} \tag{13.29a}$$

$$\Delta U = \tfrac{3}{2}(p_2 - p_1)V_1 \qquad Q = \tfrac{3}{2}(p_2 - p_1)V_1 \qquad W = 0 \tag{13.29b}$$

From this we can see how such a process could be realized: we simply keep the gas in a vessel of constant volume and heat it. All of the heat energy absorbed by the gas goes directly into its internal energy, with the result that its temperature and hence pressure rise.

Processes that, like the vertical leg of this cycle, keep volume constant are called isochoric. Because there is no change in volume, no work is ever done during an isochoric process:

$$W_{isochoric} = 0$$

The curved leg of the cycle of fig. (13.1) represents an isothermal expansion, that is, an expansion of the gas at constant temperature. By $pV = NkT$, constant temperature means that

$$pV = NkT = \text{constant}$$

throughout an isothermal process. Isothermal curves in the pV plane are therefore of the form

$$p \propto \frac{1}{V}$$

To realize an isothermal process, we need a means of both holding the temperature constant and, while doing so, adjusting the volume of the gas. We could, for example, fit the vessel containing the gas with a movable piston and then immerse the vessel in a bath of the same temperature, making changes in the volume of the gas slowly enough that the gas would remain in thermal equilibrium with the bath: as long as the bath were large enough to be considered infinite, its temperature would not change as it shed heat

into or absorbed heat from the gas. This could be rather clumsy to arrange in practice, but in principle we could set up such a bath.

For the cycle of fig. (13.1), we have already worked out that the temperature at the end of the vertical leg of the cycle of fig. (13.1) is p_2V_1/Nk, so that

$$T = \frac{p_2V_1}{Nk}$$

throughout the isothermal leg and thus

$$pV = NkT = Nk\,\frac{p_2V_1}{Nk} = p_2V_1$$

This means that in particular at the final point (p_1, V_2) of this leg we must have

$$p_1V_2 = p_2V_1$$

In other words, our pressures p_1 and p_2 and volumes V_1 and V_2 cannot have just any old values; once we have chosen the values of three of them, $p_1V_2 = p_2V_1$ will give us the value of the fourth. Having, for example, chosen the values of p_1, p_2, and V_1, we must have $V_2 = p_2V_1/p_1$. Physically, this corresponds to the volume at which the pressure returns to its initial value p_1 being determined by the isothermal expansion that returns us to that pressure. Graphically, it corresponds to there being a definite point at which the horizontal line representing pressure p_1 intersects the isothermal curve.

Since the temperature is constant, there is no change in the internal energy $U = \frac{3}{2}NkT$ of the gas. The gas does, however, do nonzero work as it expands: since $pV = p_2V_1$ throughout the expansion, $p = p_2V_1/V$ and

$$W = \int_{V_1}^{V_2} p\,dV = \int_{V_1}^{V_2} \frac{p_2V_1}{V}\,dV = p_2V_1\Big[\ln V\Big]_{V_1}^{V_2} = p_2V_1 \ln\frac{V_2}{V_1}$$

The first law then gives us

$$\begin{aligned}\Delta U &= Q - W\\ 0 &= Q - p_2V_1\ln\frac{V_2}{V_1}\\ Q &= p_2V_1\ln\frac{V_2}{V_1}\end{aligned}$$

To summarize, for the isothermal expansion we have

$$T = \frac{p_2V_1}{Nk}\ \text{(constant)} \qquad p = p_2 \to p_1 \qquad V = V_1 \to V_2 \tag{13.30a}$$

$$\Delta U = 0 \qquad Q = p_2V_1\ln\frac{V_2}{V_1} \qquad W = p_2V_1\ln\frac{V_2}{V_1} \tag{13.30b}$$

along with the condition that

$$p_1 V_2 = p_2 V_1 \tag{13.31}$$

From this we see that as the gas expands from volume V_1 to volume V_2 at constant temperature, the pressure drops from p_2 back to p_1. Since the gas does work as it expands, it must absorb an equal amount of heat energy in order to keep its internal energy and hence its temperature constant.

On the final leg of the cycle of fig. (13.1), the gas is returned from volume V_2 to its original volume V_1 by compressing it at constant pressure p_1. Processes that keep the pressure constant are called isobaric. The temperature on this isobaric leg will go from the value $p_2 V_1/Nk$ that it had along the isothermal expansion back to its original value $p_1 V_1/Nk$. The change in the internal energy of the gas during this leg is therefore

$$\Delta U = \tfrac{3}{2} Nk \Delta T = \tfrac{3}{2} Nk \left(\frac{p_1 V_1}{Nk} - \frac{p_2 V_1}{Nk} \right) = -\tfrac{3}{2}(p_2 - p_1) V_1$$

Since the pressure is constant at value p_1, the work done is

$$W = \int_{V_2}^{V_1} p \, dV = \int_{V_2}^{V_1} p_1 \, dV = p_1 \Big[V \Big]_{V_2}^{V_1} = -p_1 (V_2 - V_1)$$

Consequently the heat energy absorbed by the gas is, by the first law,

$$\begin{aligned} \Delta U &= Q - W \\ -\tfrac{3}{2}(p_2 - p_1) V_1 &= Q + p_1 (V_2 - V_1) \\ Q &= -p_1 (V_2 - V_1) - \tfrac{3}{2}(p_2 - p_1) V_1 \end{aligned}$$

Note that since $p_1 < p_2$ and $V_1 < V_2$, ΔU, Q, and W are all negative.

To summarize, for the isobaric compression we have

$$p = p_1 \text{ (constant)} \qquad V = V_2 \to V_1 \qquad T = \frac{p_2 V_1}{Nk} \to \frac{p_1 V_1}{Nk} \tag{13.32a}$$

$$\Delta U = -\tfrac{3}{2}(p_2 - p_1) V_1 \qquad W = -p_1 (V_2 - V_1)$$

$$Q = -p_1 (V_2 - V_1) - \tfrac{3}{2}(p_2 - p_1) V_1 \tag{13.32b}$$

The negative value for W corresponds to our having to do work *on* the gas to compress it. To keep the pressure constant as the gas is compressed into a smaller volume, the temperature and hence the internal energy of the gas must be reduced. Since the work that we do on the gas would otherwise increase its internal energy, the gas must shed heat energy equal to the sum of the work we do on it and the required reduction in internal energy.

To realize an isobaric process in practice is easy, at least at one atmosphere of pressure: you simply fit the vessel containing the gas with a freely movable

piston, the far side of which is open to the atmosphere, and then heat or cool the gas. Whenever the pressure of the gas in the vessel grows infinitesimally higher or drops infinitesimally lower than atmospheric pressure, the pressure difference will cause the piston to move outward or inward until the pressure equalizes at one atmosphere. Almost all household thermal processes are isobaric precisely because they are carried out in the open atmosphere. If a constant pressure other than one atmosphere is desired, this requires a bit more engineering but is still not difficult: weights can be used to press or pull on the outer side of the piston.[28] The weight of a mass m pushing or pulling on a piston of cross-sectional area A will produce a pressure of

$$1\text{ atm} \pm \frac{mg}{A}$$

respectively.

For the cycle overall, we see from eqq. (13.29b), (13.30b), and (13.32b) that the net change in the internal energy does indeed vanish:

$$\Delta U_{cycle} = \tfrac{3}{2}(p_2 - p_1)V_1 + 0 - \tfrac{3}{2}(p_2 - p_1)V_1 = 0 \quad \checkmark$$

But even though the gas returns to the same start from which it started out, the net work done and the net heat energy absorbed are nonzero:

$$\begin{aligned} W_{net} &= 0 + p_2 V_1 \ln\frac{V_2}{V_1} - p_1(V_2 - V_1) \\ &= p_2 V_1 \ln\frac{V_2}{V_1} - p_1(V_2 - V_1) \qquad (13.33a) \\ Q_{net} &= \tfrac{3}{2}(p_2 - p_1)V_1 + p_2 V_1 \ln\frac{V_2}{V_1} + \Big(-p_1(V_2 - V_1) - \tfrac{3}{2}(p_2 - p_1)V_1\Big) \\ &= p_2 V_1 \ln\frac{V_2}{V_1} - p_1(V_2 - V_1) \qquad (13.33b) \end{aligned}$$

Q and W are therefore not state variables: if they depended only on the state of the system, they would return to their initial values as the system returned to its initial state and the net Q and W would therefore vanish. Note also that $W_{net} = Q_{net}$: since ΔU vanishes around any cycle, the first law dictates that the net work done around a cycle must equal the net heat energy absorbed:

$$\begin{aligned} \Delta U_{cycle} &= Q_{net} - W_{net} \\ 0 &= Q_{net} - W_{net} \\ W_{net} &= Q_{net} \end{aligned}$$

[28] For those of you who are "mechanically challenged," to pull you can tie a cord to the piston, sling it over a pulley, and hang weights from the far end.

In this example we went around the cycle clockwise: the gas was heated isochorically, then underwent an isothermal expansion, then was compressed isobarically back to its original state. No work was done during the isochoric heating, the gas did positive work

$$p_2 V_1 \ln \frac{V_2}{V_1}$$

during the isothermal expansion, and during the isobaric compression we did work to compress the gas, with the result that the work done by the gas was negative:

$$-p_1(V_2 - V_1)$$

If we think of the work done by the gas geometrically, as the area under the curves for these three legs in the pV plane, then the area under the line for the isobaric compression will be negative because we are going "backward" to a smaller volume. As you can see from fig. (13.1), the positive area under the isothermal expansion curve is clearly greater than the negative area under the isobaric line, so that the net work done by the gas is positive. While it is not as immediately obvious, we can also see this algebraically from eq. (13.33a):

$$\begin{aligned} W_{net} &= p_2 V_1 \ln \frac{V_2}{V_1} - p_1(V_2 - V_1) \\ &= p_2 V_1 \ln \frac{V_2}{V_1} - p_1 V_2 + p_1 V_1 \end{aligned}$$

If we use condition (13.31),

$$p_1 V_2 = p_2 V_1$$

this W_{net} can be re-expressed as

$$\begin{aligned} W_{net} &= p_2 V_1 \ln \frac{V_2}{V_1} - p_2 V_1 + p_1 V_1 \\ &= p_2 V_1 \left(\ln \frac{V_2}{V_1} - 1 \right) + p_1 V_1 \end{aligned}$$

which is unambiguously positive for $V_2 > V_1$. Since $W_{net} = Q_{net}$, Q_{net} is also positive. So when we go around the cycle clockwise, the net effect is that the gas absorbs a net amount of heat energy and does an equal net amount of work. In other words, this cycle converts heat into work. Cycles that accomplish this are called heat engines, and we will be saying a great deal more about them in §13.7.

We could also have gone around the cycle of fig. (13.1) counterclockwise. Before you start rioting, note that this does not mean repeating the above analysis in its entirety. It doesn't even mean repeating it in part. We need simply recognize that the effect of reversing the initial and final states of

each leg will be a reversal of the signs on all the ΔU's, Q's, and W's. Run backward, the cycle will therefore do a negative net amount of work, meaning that we will have to do work to put the gas through the cycle. And when we reverse the signs on the Q's in eqq. (13.29b), (13.30b), and (13.32b), we see that

$$\begin{aligned} Q_{isochoric} &= -\tfrac{3}{2}(p_2 - p_1)V_1 < 0 \\ Q_{isothermal} &= -p_2 V_1 \ln \frac{V_2}{V_1} < 0 \\ Q_{isobaric} &= p_1(V_2 - V_1) + \tfrac{3}{2}(p_2 - p_1)V_1 > 0 \end{aligned}$$

The gas thus absorbs heat during the isobaric leg (which is now an expansion rather than a compression) and sheds it during the isochoric and isothermal legs. Run counterclockwise, our cycle therefore functions as a refrigerator: we have to do work to run the cycle, but the payoff is that during the isobaric leg the gas sucks up heat, removing it from what would constitute the interior of the refrigerator. We will also be saying a great deal more about refrigerators in §13.7.

13.6.2 A Mercifully Short Example

As a further example of how Q and W depend not only on the initial and final state of the system but also on the process that brings about the change, consider two of the infinitely many possible paths that would take a sample of an ideal monatomic gas from the initial state S_i to the final state S_f shown in fig. (13.2): the blue path ab and the red path cd.

Along the path ab, the gas expands isobarically on leg a and then is heated isochorically on leg b. During the isobaric expansion, the pressure is

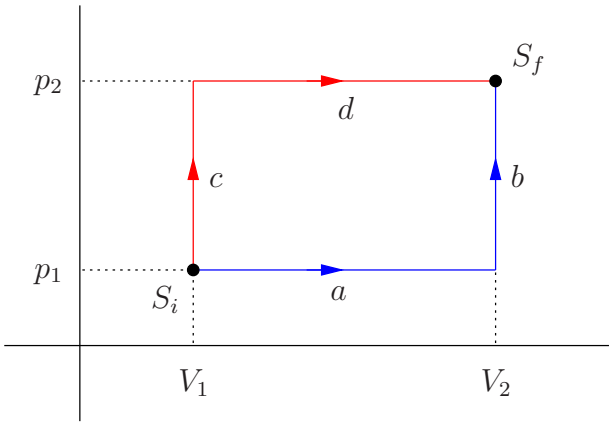


Figure 13.2: Two Paths Between the Same Two States

constant at p_1, so that, if we use $pV = NkT$ and the first law,

$$\begin{aligned}
W_a &= \int_{V_1}^{V_2} p\,dV = p_1 \int_{V_1}^{V_2} dV = p_1(V_2 - V_1) \\
(\Delta U)_a &= \Delta\Big(\tfrac{3}{2}NkT\Big) = \Delta\Big(\tfrac{3}{2}pV\Big) = \tfrac{3}{2}p_1\Delta V = \tfrac{3}{2}p_1(V_2 - V_1) \\
(\Delta U)_a &= Q_a - W_a \\
Q_a &= (\Delta U)_a + W_a = \tfrac{3}{2}p_1(V_2 - V_1) + p_1(V_2 - V_1) = \tfrac{5}{2}p_1(V_2 - V_1)
\end{aligned}$$

During the isochoric heating, no work is done because there is no change in volume, and we have

$$\begin{aligned}
W_b &= 0 \\
(\Delta U)_b &= \Delta\Big(\tfrac{3}{2}NkT\Big) = \Delta\Big(\tfrac{3}{2}pV\Big) = \tfrac{3}{2}(\Delta p)V_2 = \tfrac{3}{2}(p_2 - p_1)V_2 \\
Q_b &= (\Delta U)_b + W_b = \tfrac{3}{2}(p_2 - p_1)V_2 + 0 = \tfrac{3}{2}(p_2 - p_1)V_2
\end{aligned}$$

The net changes over path ab are thus

$$\begin{aligned}
W_{ab} &= W_a + W_b \\
&= p_1(V_2 - V_1) + 0 \\
&= p_1(V_2 - V_1) \\
(\Delta U)_{ab} &= (\Delta U)_a + (\Delta U)_b \\
&= \tfrac{3}{2}p_1(V_2 - V_1) + \tfrac{3}{2}(p_2 - p_1)V_2 \\
&= \tfrac{3}{2}(p_2V_2 - p_1V_1) \\
Q_{ab} &= Q_a + Q_b \\
&= \tfrac{5}{2}p_1(V_2 - V_1) + \tfrac{3}{2}(p_2 - p_1)V_2
\end{aligned}$$

An almost identical analysis yields for leg c

$$\begin{aligned}
W_c &= 0 \\
(\Delta U)_c &= \Delta\Big(\tfrac{3}{2}NkT\Big) = \Delta\Big(\tfrac{3}{2}pV\Big) = \tfrac{3}{2}(\Delta p)V_1 = \tfrac{3}{2}(p_2 - p_1)V_1 \\
Q_c &= (\Delta U)_c + W_c = \tfrac{3}{2}(p_2 - p_1)V_1 + 0 = \tfrac{3}{2}(p_2 - p_1)V_1
\end{aligned}$$

and for leg d

$$\begin{aligned}
W_d &= \int_{V_1}^{V_2} p\,dV = p_2 \int_{V_1}^{V_2} dV = p_2(V_2 - V_1) \\
(\Delta U)_d &= \Delta\Big(\tfrac{3}{2}NkT\Big) = \Delta\Big(\tfrac{3}{2}pV\Big) = \tfrac{3}{2}p_2\Delta V = \tfrac{3}{2}p_2(V_2 - V_1) \\
(\Delta U)_d &= Q_d - W_d
\end{aligned}$$

$$Q_d = (\Delta U)_d + W_d = \tfrac{3}{2}p_2(V_2 - V_1) + p_2(V_2 - V_1) = \tfrac{5}{2}p_2(V_2 - V_1)$$

and therefore for the path cd

$$\begin{aligned} W_{cd} &= W_c + W_d \\ &= 0 + p_2(V_2 - V_1) \\ &= p_2(V_2 - V_1) \\ (\Delta U)_{cd} &= (\Delta U)_c + (\Delta U)_d \\ &= \tfrac{3}{2}(p_2 - p_1)V_1 + \tfrac{3}{2}p_2(V_2 - V_1) \\ &= \tfrac{3}{2}(p_2V_2 - p_1V_1) \\ Q_{cd} &= Q_c + Q_d \\ &= \tfrac{3}{2}(p_2 - p_1)V_1 + \tfrac{5}{2}p_2(V_2 - V_1) \end{aligned}$$

From this we see that the net heat energy absorbed and the net work done both differ along the paths ab and cd: $Q_{ab} \neq Q_{cd}$ and $W_{ab} \neq W_{cd}$. As we would expect because of the greater area under the curve, more work is done along path cd than along path ab: $p_2(V_2 - V_1) > p_1(V_2 - V_1)$. But since U is a state variable, ΔU depends only on the initial state S_i and final state S_f and not on the path from S_i to S_f: ΔU is the same for path cd as it is for path ab and would be the same for any path from S_i to S_f. And since ΔU is the same along any two paths P_1 and P_2 between the same initial and final states, the first law quite generally gives us

$$\begin{aligned} \Big((\Delta U)_{P_2} - (\Delta U)_{P_1}\Big) &= (Q_{P_2} - W_{P_2}) - (Q_{P_1} - W_{P_1}) \\ 0 &= (Q_{P_2} - Q_{P_1}) - (W_{P_2} - W_{P_1}) \\ Q_{P_2} - Q_{P_1} &= W_{P_2} - W_{P_1} \end{aligned}$$

In other words, although the Q and W along one path in general differ from the Q and W along another path, the difference in the Q's along two paths will always equal the difference in the W's along those paths. In our present example, this means that the difference in heat energy absorbed along paths ab and cd must equal the difference in the work done along them, with more heat energy being absorbed along path cd than along path ab, as we can verify explicitly:

$$\begin{aligned} W_{cd} - W_{ab} &= p_2(V_2 - V_1) - p_1(V_2 - V_1) \\ &= (p_2 - p_1)(V_2 - V_1) \\ Q_{cd} - Q_{ab} &= \Big(\tfrac{3}{2}(p_2 - p_1)V_1 + \tfrac{5}{2}p_2(V_2 - V_1)\Big) \\ &\quad - \Big(\tfrac{5}{2}p_1(V_2 - V_1) + \tfrac{3}{2}(p_2 - p_1)V_2\Big) \end{aligned}$$

$$\begin{aligned}
&= (-\tfrac{3}{2} + \tfrac{5}{2})p_1V_1 + (\tfrac{3}{2} - \tfrac{5}{2})p_2V_1 \\
&\qquad + (-\tfrac{5}{2} + \tfrac{3}{2})p_1V_2 + (\tfrac{5}{2} - \tfrac{3}{2})p_2V_2 \\
&= p_1V_1 - p_2V_1 - p_1V_2 + p_2V_2 \\
&= (p_2 - p_1)(V_2 - V_1) \quad \checkmark
\end{aligned}$$

13.6.3 Adiabatic Processes

The Greek root of *adiabatic* is ἀδιαβάτικος: the α-privative for negation, plus -δια- (through), plus -βατικος from the verb βαίνω (to go).[29] The root sense is thus *not going through* or *across*, which in the context of thermodynamics means that no heat energy crosses the boundary between the system and its environment. Although the Greeks themselves never used the word in the thermodynamic sense — or in any other sense, for that matter. But in modern science it's considered really cool to use obscure Greek and Latin terms for stuff.

Anyway, an adiabatic process is by definition one for which $Q = 0$. To realize an adiabatic process in practice, one has therefore simply to ensure that the system neither absorbs nor sheds any heat. This can be accomplished by thermally insulating the system to prevent heat exchange with its environment or by making the change associated with the adiabatic process so quickly that there isn't time for significant heat exchange with the environment.

By the first law, an immediate consequence of the vanishing of Q for adiabatic processes is that

$$\Delta U = Q - W = 0 - W = -W$$

In other words, since the work done *on* the system is the negative of the work done *by* the system, for an adiabatic process the change in the internal energy of the system equals the work done on the system:

$$\Delta U = -W_{by\ system} = W_{on\ system}$$

Determining the relations among p, V, and T is, however, a bit more involved. We will first take a detour through some properties of polyatomic molecules for isochoric and isobaric processes. Since the work done by a gas is $dW = p\,dV$, we can express the first law as

$$dU = dQ - dW = dQ - p\,dV$$

and thus

$$dQ = dU + p\,dV \tag{13.34}$$

[29] Isn't your life just so much richer now for knowing that?

Eq. (13.34) is valid for any process to which a sample of gas is subjected.

For an ideal monatomic gas, we saw in §13.5 that $U = \frac{3}{2}NkT$ as a result of the freedom of the molecules to move along any of the three spatial directions: for each of the N molecules there was $\frac{1}{2}kT$ of energy for each of the three translational degrees of freedom. A polyatomic molecule will generally also have rotational and vibrational degrees of freedom, so that it has $D > 3$ of degrees of freedom and its internal energy is consequently

$$U = N \cdot \tfrac{1}{2}kT \cdot D = \tfrac{1}{2}DNkT \tag{13.35}$$

For such a molecule, eq. (13.34) becomes

$$\begin{aligned} dQ &= d\left(\tfrac{1}{2}DNkT\right) + p\,dV \\ &= \tfrac{1}{2}DNk\,dT + p\,dV \end{aligned} \tag{13.36}$$

The arguments by which we derived the ideal-gas law from the kinematics of molecular collisions with the walls of the vessel containing the gas will, however, be the same for a polyatomic as they were for a monatomic ideal gas, so that polyatomic ideal gases still obey $pV = NkT$.

By definition, the heat energy dQ needed to raise the temperature of a sample of gas isochorically by dT is[30]

$$dQ = C_V\,dT$$

where C_V is the heat capacity at constant volume. For an isochoric process, the first law thus becomes

$$dU = dQ - dW = C_V\,dT - 0 = C_V\,dT$$

Similarly the heat energy needed to raise the temperature of a sample of gas isobarically by dT is defined to be

$$dQ = C_p\,dT$$

where C_p is the heat capacity at constant pressure. Using $dQ = C_V\,dT$ and $dV = 0$ in eq. (13.36), we see that for an isochoric process

$$\begin{aligned} dQ &= \tfrac{1}{2}DNk\,dT + p\,dV \\ C_V\,dT &= \tfrac{1}{2}DNk\,dT + 0 \end{aligned}$$

and hence

$$C_V = \tfrac{1}{2}DNk \tag{13.37}$$

[30] As you can see by comparing eq. (13.7) in the differential form $dQ = mc\,dT$ with $Q = C_V\,dT$, the heat capacity C_V is just the product of the sample's mass and specific heat, mc, with the V subscript indicating that the volume is held constant.

And since the pressure is constant for an isobaric process, the ideal-gas law gives

$$p\,dV = d(pV) = d(NkT) = Nk\,dT$$

Using this and $dQ = C_p\,dT$ in eq. (13.36), we see that for an isobaric process

$$\begin{aligned} dQ &= \tfrac{1}{2}DNk\,dT + p\,dV \\ C_p\,dT &= \tfrac{1}{2}DNk\,dT + Nk\,dT \\ &= \left(\tfrac{1}{2}D + 1\right) Nk\,dT \end{aligned}$$

and hence

$$C_p = \left(\tfrac{1}{2}D + 1\right) Nk \tag{13.38}$$

Thus C_p is greater than C_V because of the work done during the expansion of the gas when the pressure rather then the volume is held constant. The number D of degrees of freedom of the gas molecules can in fact be determined by measuring C_p and C_V:

$$\frac{C_p}{C_V} = \frac{\left(\tfrac{1}{2}D + 1\right) Nk}{\tfrac{1}{2}DNk} = \frac{\tfrac{1}{2}D + 1}{\tfrac{1}{2}D}$$

Knowing the ratio C_p/C_V, one can then solve for D. This ratio shows up frequently enough in our present and other contexts that it even has its own symbol, γ:

$$\gamma = \frac{C_p}{C_V} = \frac{\tfrac{1}{2}D + 1}{\tfrac{1}{2}D} \tag{13.39}$$

For a monatomic ideal gas, $D = 3$ and $\gamma = \frac{5}{3}$. For gases with more degrees of freedom, γ is smaller, and in the limit $D \to \infty$, $\gamma \to 1$: when there are many degrees of freedom, most of the heat added to the system is sucked up by the extra rotational and vibrational modes, so that the additional energy required with C_p for the expansion of the gas becomes relatively insignificant.[31]

[31] We don't want to go too far off on a tangent, but it is worth noting that just because a system has an extra degree of freedom doesn't necessarily mean that that degree of freedom contributes to the system's internal energy: in quantum theory, degrees of freedom may, so to speak, be frozen until the temperature reaches a certain threshold. Angular momentum L is quantized, for example, so that when a molecule's moment of inertia I along an axis is small, the rotational kinetic energy $E_{rot} = \frac{1}{2}I\omega^2 = L^2/2I$ corresponding to the lowest nonzero value of L is large enough that the relative probability $e^{-E_{rot}/kT}$ of being in that state is vanishingly small at all except very high temperatures T. This is why we didn't have to worry about the rotational modes of monatomic molecules: while not point-like, they are small enough that their rotational degrees of freedom are frozen at ordinary temperatures. As gases with frozen degrees of freedom are heated, one sees the effective value of D grow in steps as the temperature crosses the thresholds for exciting these frozen degrees.

We now return to adiabatic processes. Since by definition $dQ = 0$ for an adiabatic process, eq. (13.36) reduces to

$$\begin{aligned} dQ &= \tfrac{1}{2}DNk\,dT + p\,dV \\ 0 &= \tfrac{1}{2}DNk\,dT + p\,dV \end{aligned} \tag{13.40}$$

Recall now that because of the equation of state $pV = NkT$, only two of p, T, and V are independent. If we choose T and V as the independent variables, then $p = NkT/V$ and eq. (13.40) becomes

$$0 = \tfrac{1}{2}DNk\,dT + \frac{NkT}{V}\,dV$$

or, if we move the $\frac{1}{2}DNk\,dT$ to the left side and divide both sides by NkT to separate variables,

$$-\tfrac{1}{2}D\,\frac{dT}{T} = \frac{dV}{V}$$

We can now integrate this between two states (V_1, T_1) and (V_2, T_2) during an adiabatic process:

$$\begin{aligned} -\tfrac{1}{2}D\int_{T_1}^{T_2}\frac{dT}{T} &= \int_{V_1}^{V_2}\frac{dV}{V} \\ -\tfrac{1}{2}D\Big[\ln T\Big]_{T_1}^{T_2} &= \Big[\ln V\Big]_{V_1}^{V_2} \\ -\tfrac{1}{2}D\ln\frac{T_2}{T_1} &= \ln\frac{V_2}{V_1} \\ \left(\frac{T_2}{T_1}\right)^{-\frac{1}{2}D} &= \frac{V_2}{V_1} \\ \left(\frac{T_1}{T_2}\right)^{\frac{1}{2}D} &= \frac{V_2}{V_1}- \\ V_1T_1^{\frac{1}{2}D} &= V_2T_2^{\frac{1}{2}D} \end{aligned}$$

Since this relation holds between any two states (V_1, T_1) and (V_2, T_2) during our adiabatic process, we must in other words have

$$VT^{\frac{1}{2}D} = \text{const} \tag{13.41}$$

While fine the way it is written, this relation is conventionally expressed in terms of γ rather than D. Inverting eq. (13.39) yields

$$\gamma = \frac{\frac{1}{2}D + 1}{\frac{1}{2}D}$$

$$\begin{aligned} \tfrac{1}{2}D\gamma &= \tfrac{1}{2}D + 1 \\ \tfrac{1}{2}D(\gamma - 1) &= 1 \\ \tfrac{1}{2}D &= \frac{1}{\gamma - 1} \end{aligned} \tag{13.42}$$

so that eq. (13.41) becomes

$$VT^{1/(\gamma-1)} = \text{const} \tag{13.43}$$

or, if we raise both sides to the power $\gamma - 1$,[32]

$$TV^{\gamma-1} = \text{const} \tag{13.44}$$

Eq. (13.44) relates T and V during an adiabatic process. To obtain a relation between p and V, we can simply use the equation of state $pV = NkT$ in the form $T = pV/Nk$ in eq. (13.44):

$$\begin{aligned} \frac{pV}{Nk}\, V^{\gamma-1} &= \text{const} \\ \frac{pV^{\gamma}}{Nk} &= \text{const} \end{aligned}$$

or, if we multiply both sides by the constant Nk,[33]

$$pV^{\gamma} = \text{const} \tag{13.45}$$

Alternatively, we could start again from eq. (13.40), this time taking p and V as the independent variables and therefore using

$$Nk\, dT = d(NkT) = d(pV) = dp\, V + p\, dV$$

in eq. (13.40):

$$\begin{aligned} 0 &= \tfrac{1}{2}DNk\, dT + p\, dV \\ &= \tfrac{1}{2}D(dp\, V + p\, dV) + p\, dV \\ &= \tfrac{1}{2}D\, dp\, V + \left(\tfrac{1}{2}D + 1\right) p\, dV \end{aligned}$$

Separating variables now gives

$$\frac{dp}{p} = \frac{\frac{1}{2}D + 1}{\frac{1}{2}D}\,\frac{dV}{V} = \gamma\,\frac{dV}{V}$$

[32] We are not, of course, saying that the constants on the right-hand sides of eqq. (13.43) and (13.44) are the same. Just in case you were getting your knickers in a twist over that one.

[33] Again, we are not saying that the constants on the right-hand sides of eqq. (13.44) and (13.45) are the same. *Sheesh!*

which when integrated yields

$$\begin{aligned}\ln\frac{p_2}{p_1} &= \gamma \ln\frac{V_2}{V_1} \\ &= \ln\left(\frac{V_2}{V_1}\right)^{\gamma}\end{aligned}$$

Yada, yada, yada.

We leave it to you to work out the relation between p and T if you're interested.

13.6.4 Some General Observations

Because an ideal gas obeys the equation of state $pV = NkT$, only two of the state variables p, V, and T are independent. If we choose p and V as the independent variables, then there is a one-to-one correspondence between the states of the gas and points in the pV plane. A process is any operation or action that changes the state of a system. Graphically, continuous processes correspond to curves in the pV plane. At each point in time during a process — or, graphically, at each point along the corresponding curve in the pV plane — , the equation of state $pV = NkT$ holds, and in addition the changes that occur during the process obey the first law

$$dU = dQ - dW$$

For an ideal gas with D degrees of freedom, the internal energy is given by

$$U = \tfrac{1}{2}DNkT$$

A system can be brought from a given initial to a given final state by an infinite variety of processes: each of the infinitely many distinct curves connecting the initial and final state in the pV plane corresponds to a distinct process. Of principal importance, however, are

Isothermal processes, for which the temperature remains constant. For an isothermal process, $\Delta U = 0$.

Isothermal processes are most easily realized by keeping the system in a bath at the desired temperature and making any changes to the volume and pressure gradually enough that the system remains in thermal equilibrium with the bath. The bath must be large enough that supplying heat to or absorbing heat from the system does not alter the bath's temperature — in principle, infinite; in practice, large enough that any changes in its temperature are negligibly small.

Isobaric processes, for which the pressure remains constant.

Isobaric processes are most easily realized by means of containers fitted with freely movable pistons: the piston will move to eliminate any infinitesimal pressure difference that arises. Weights can be used as described on p.692 when a constant pressure differing from atmospheric pressure is required.

Isochoric processes, for which the volume remains constant. Because there is no change in volume, an isochoric process does no work: $W = 0$.

Isochoric processes are easily realized by confining the system within a container of fixed volume.

Adiabatic processes, for which $Q = 0$, that is, during which the system does not exchange any heat with its environment. For adiabatic processes involving ideal gases, pV^γ and $TV^{\gamma-1}$ are constant, where γ is the ratio of heat capacity at constant pressure to heat capacity at constant volume and is given, for molecules with D degrees of freedom, by

$$\gamma = \frac{C_p}{C_V} = \frac{\frac{1}{2}D + 1}{\frac{1}{2}D}$$

Adiabatic processes can be realized either by thermally insulating the system or by making the changes in volume and pressure so quickly that there isn't time for significant heat exchange with the environment.

Because doing calculations for nonequilibrium states is prohibitively difficult, we have tacitly assumed in everything we have done so far that the system is always in an equilibrium state and that any changes in its state are made gradually through a succession of equilibrium states, each differing only infinitesimally from its predecessor. Changes wrought through such successions of equilibrium states are called quasistatic. For an isothermal process, for example, this means that the changes in pressure and volume must be made gradually enough that the system has time to absorb or shed the heat necessary to keep its temperature constant both temporally and spatially: the temperature must at any given instant be uniform throughout the system. For an adiabatic process, it means that the change in volume and pressure must be gradual enough that the temperature, while changing with time, likewise remains uniform throughout the system at any given instant. It is possible to make an adiabatic change so quickly that this condition is not met: the compression of a piston, for example, can be so rapid that its effects do not have time to propagate beyond the molecules adjacent to the piston and distribute themselves uniformly throughout the gas. In practice, however, this is not usually a concern. At any rate, if you're capable of compressing gases at supersonic speeds, you could make a good living as a circus performer. Or at least it would make an entertaining party trick.

13.7 Heat Engines, & Refrigerators

A heat engine is a device for converting heat energy — as opposed to the mechanical potential energy stored in a compressed spring or the electrical potential energy stored in a battery, etc. — into useful mechanical work, such as driving a turbine at a power plant. There is no limit on the variety of schemes by which you can accomplish this; any cycle that does positive net work, and thus any closed curve traversed clockwise in the pV plane, constitutes a heat engine. The cycle discussed in §13.6.1, for example, qualifies.

To realize a cyclic heat engine, all you need is a sample of gas enclosed in a cylinder fitted with a movable piston: you can heat or cool the gas through the walls of the cylinder, and you can adjust the volume and pressure by means of the piston. You could, for example, start with such an apparatus at room temperature and place it in a vat of boiling water. As the temperature of the gas inside rose, it would expand outward against the piston, thereby doing work for you.[34] When the expansion petered out, you would simply yank the cylinder out of the hot water and leave it in the open room until it returned to its original state, whereupon you would repeat the process. You wouldn't win any engineering awards, but heat would be converted into work.

As we will see in more detail in §13.7.2, a refrigerator is just a heat engine run backward: instead of putting heat into the engine and getting work out of it, you put work into running the refrigerator and it removes heat for you. You could, for example, cool your home in the summer with the same apparatus by adiabatically expanding the gas until its temperature dropped below room temperature. You would then wait while the gas was returned to room temperature by heat absorbed from the room, at which point you could take the apparatus outside, adiabatically compress until it is hotter than the great outdoors, wait for it to cool off to the ambient temperature, and then bring it back inside and repeat the process. Not a scheme likely to attract investment, but in principle this constitutes a refrigerator: by virtue of the work you do on the gas, heat is absorbed from the room and transferred to the great outdoors.

13.7.1 The Carnot Cycle

The Carnot cycle was conceived by Nicolas Léonard Sadi Carnot[35] in a fit of genius in 1824. It is not a cycle most people would think of, and in fact

[34] Just how you would convert this outward push of the piston into useful mechanical work is an engineering detail that need not concern us; in principle work is being done for you.

[35] If you haven't already guessed, Carnot was a French dude, so the pronunciation is Kär-nō′, not Kär′-nŏt.

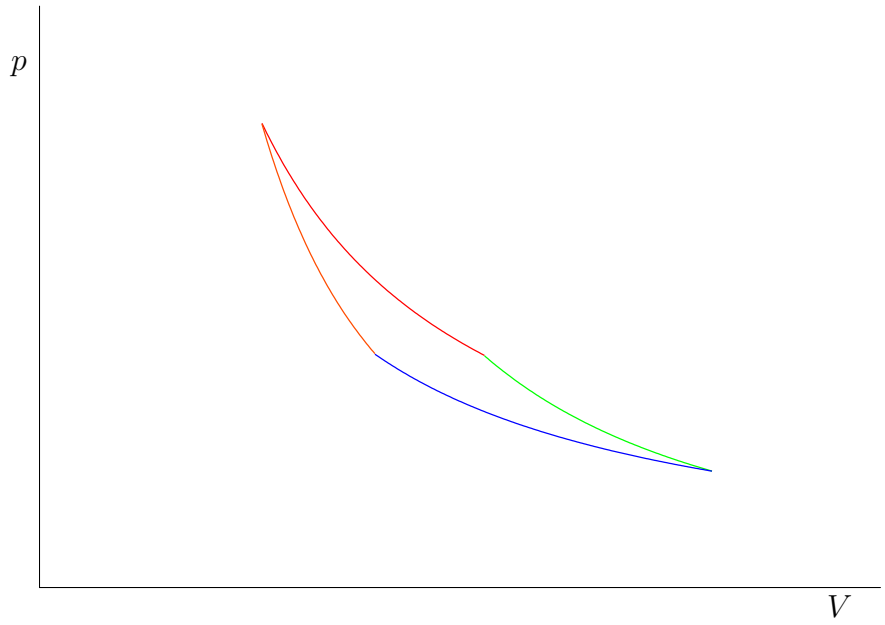


Figure 13.3: Plot of an Actual Carnot Cycle.
Viewer Discretion is Advised.

it may at first seem to have been conceived in a fit of a very different sort involving crystal meth, but it is of great theoretical importance for reasons that we will discuss in §13.8.

We will introduce you to the Carnot cycle in three stages: first we will describe the sequence of processes that constitutes the cycle, then we will discuss how a Carnot cycle could be realized in practice, and finally we will work through the calculations for the cycle.

And so, to get down to business: The pV plot of a Carnot cycle is shown in fig. (13.3). The first thing you will notice is that the Carnot cycle has a real attitude and, from the point of view of geometrical æsthetics, is a rather gnarly beast. You'd have a hard time packing it in a suitcase.

Of the four legs in the cycle, two (the red and the blue in fig. (13.3)) are isothermal and two (the green and the yucky sort of orange) are adiabatic. The cycle takes place between a high temperature T_H and a low (and hence relatively cold) temperature T_C. If we start at the upper left "corner" and go around clockwise, the cycle consists of

> An **isothermal expansion** at temperature T_H (the red line in fig. (13.3)). As we saw in the example of §13.6.1, during such an isothermal expansion heat is absorbed by the system, the system does work, and the pressure drops. When the gas has expanded to some arbitrarily chosen volume that is to our liking, we stop the isothermal expansion.

An **adiabatic expansion** (the green line in fig. (13.3)). As we saw in §13.6.3, during an adiabatic expansion no heat is absorbed or shed by the system, the system does work, and the pressure and temperature drop. When the temperature has dropped to T_C, we stop the adiabatic expansion.

An **isothermal compression** at temperature T_C (the blue line in fig. (13.3)). During this isothermal compression heat is shed by the system, we do work on the system, and the pressure rises. We stop the isothermal compression when the gas has been compressed to just the right volume that the system will be returned to its initial state by the final leg of the cycle.

An **adiabatic compression** (the yucky sort of orange line in fig. (13.3)). During this adiabatic compression no heat is absorbed or shed by the system, we do work on the system, and the pressure and temperature rise. When the temperature has risen to T_H, we stop the adiabatic expansion.

The temperatures T_H and T_C can be any two temperatures we want. In other words, the isothermal curves that form the "top" and "bottom" of the Carnot cycle can be any two isothermal curves. Similarly the adiabatic curves that form the "left" and "right" sides of the cycle can be any two adiabatic curves, which amounts to our being free to choose the volumes at the upper left and upper right corners of the cycle. We don't want to make too big a deal out of this right now — it will be become more clear below when we do the calculations for the Carnot cycle — , but once we have chosen the two adiabatic curves, the points at which they intersect the lower isothermal line, and thus the volumes at the lower right and lower left corners of the cycle, are determined. As noted in our description of the four legs of the Carnot cycle above, the volume at the lower right corner will be the volume corresponding to the gas having dropped to temperature T_C, and the volume at the lower left corner must be such that when the adiabatic compression has raised the system back to temperature T_H the system will be restored to its initial volume and pressure — otherwise we would not have a closed loop. When we do the calculations for the Carnot cycle, we will see that we therefore have conditions on the values of the volumes similar to eq. (13.31) in §13.6.1.

At the risk of inducing nausea, in fig. (13.4) we have plotted multiple isothermal lines in blue and multiple adiabatic lines in red: the intersections of any pair of blue isotherms and any pair of red adiabatic lines forms a funky diamond-like quadrilateral that constitutes a Carnot cycle. How many Carnot cycles can you find? [36]

[36] Those of you who actually counted will also be surprised to learn that the word *gullible* is not in the dictionary.

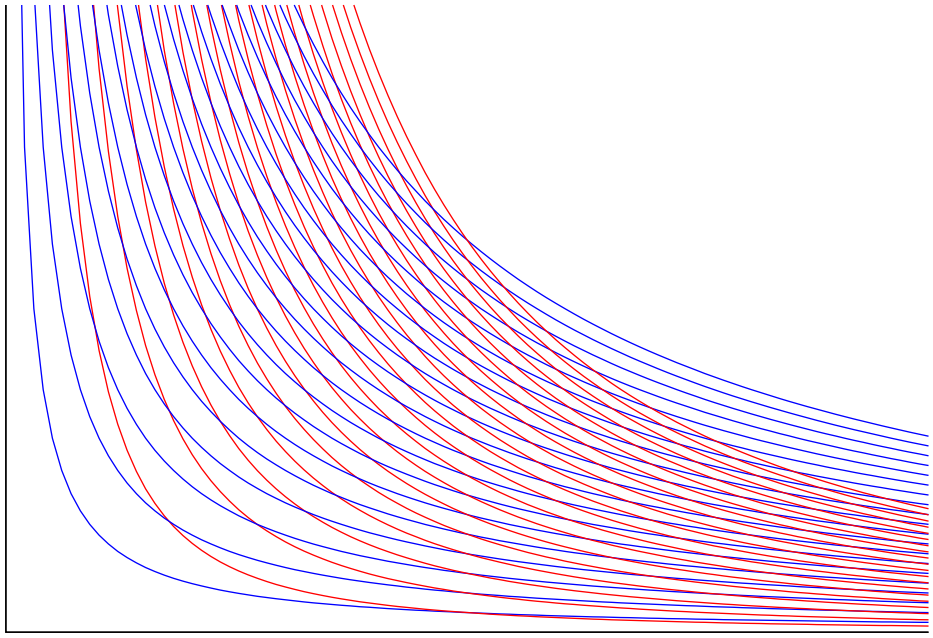


Figure 13.4: Carnot Madness

Fig. (13.5) shows how we might realize a Carnot cycle with a sample of gas enclosed in a cylinder fitted with a freely movable piston. For the isothermal legs of the cycle, we use two reservoirs, a hot reservoir at temperature T_H and a cold reservoir at temperature T_C. In principle these reservoirs should be infinite, so that their temperatures are not altered when the system absorbs heat from or sheds heat into them; in practice, they need simply be large enough that any temperature change is negligible. Alternatively, their temperatures can be actively maintained: we could, for example, use a vat of boiling water as the hot reservoir and a vat of ice water as the cold reservoir to maintain even temperatures of 100°C and 0°C.

To carry out the cycle, we perform the isothermal expansion slowly enough to ensure that the gas has time to absorb from the hot reservoir the heat necessary to keep its temperature uniformly constant at T_H. As we transfer the cylinder from the hot to the cold reservoir, heat exchange between the gas and its environment during the adiabatic expansion is then most easily prevented by performing this expansion so quickly that there isn't time for heat to flow into or out of the system. We follow this with a similarly slow isothermal compression while immersed cold reservoir and a similarly rapid adiabatic compression as we transfer the cylinder back to the hot reservoir.

If this were an English course, we would now get into a discussion of tone, irony, and symbolism in the Carnot cycle. But it's not an English course, so instead we'll calculate what actually happens during the cycle. This will

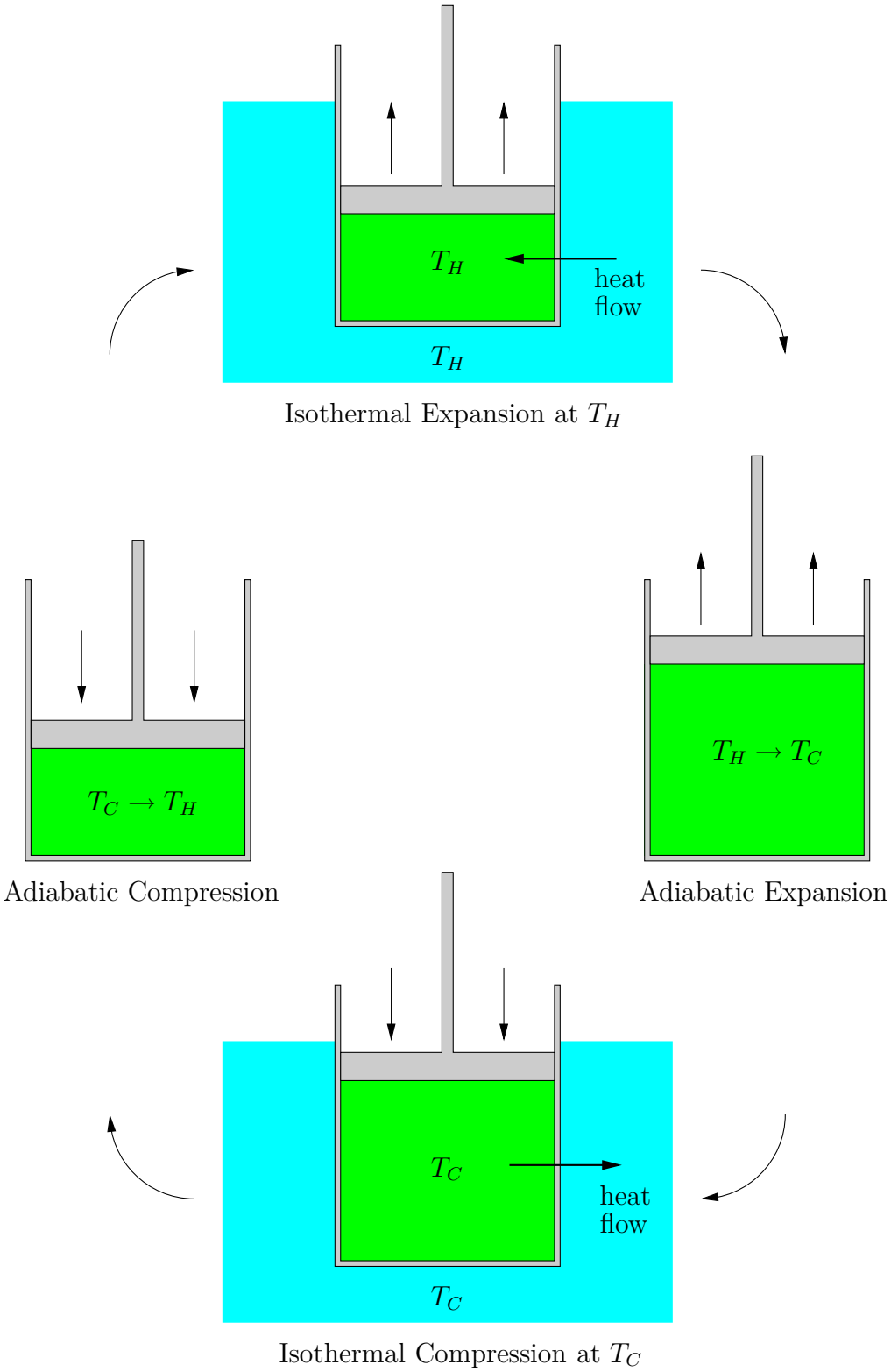


Figure 13.5: The Carnot Cycle: FR33 AmaTEuR PICS 4 U!!!

be a bit of a long haul, though not as long as the example of §13.6.1. Our concern will be with the state variables p, V, and T of the gas, and with the thermodynamic quantities U, Q, and W. As a matter of bookkeeping, let us number the vertices of the cycle clockwise, starting from the upper left, so that the vertices and legs are

Vertices:	1	upper left	Legs:	12	isothermal expansion
	2	upper right		23	adiabatic expansion
	3	lower left		34	isothermal compression
	4	lower left		41	adiabatic compression

The work W done by the gas and its changes in internal energy ΔU will be denoted accordingly: W_{12}, for example, for the work done during the isothermal expansion, and ΔU_{41} for the change in internal energy during the adiabatic compression. But since $Q = 0$ for the two adiabatic legs of the cycle, there is only the heat absorbed from the hot reservoir during the isothermal expansion and the heat shed into the cold reservoir during the isothermal compression, which we will therefore instead denote by Q_H and Q_C, respectively. (Since the system actually sheds heat during the isothermal compression, Q_C will turn out to be negative.)

We begin with the isothermal expansion. Using the equation of state $pV = NkT$ together with $dW = p\,dV$, and noting that the temperature is a constant T_H throughout the expansion, we have

$$W_{12} = \int_{V_1}^{V_2} p\,dV = \int_{V_1}^{V_2} \frac{NkT_H}{V}\,dV = NkT_H\Big[\ln V\Big]_{V_1}^{V_2} = NkT_H \ln\frac{V_2}{V_1}$$

If we further note that $\Delta U = 0$ for any isothermal process, the first law then gives us

$$\begin{aligned} \Delta U_{12} &= Q_H - W_{12} \\ 0 &= Q_H - W_{12} \\ Q_H &= W_{12} = NkT_H \ln\frac{V_2}{V_1} \end{aligned} \tag{13.46}$$

Since $V_2 > V_1$, Q_H and W_{12} are, as expected, positive: during the isothermal expansion, the system does work and absorbs heat from the hot reservoir.

To get the Q and W for the isothermal compression at temperature T_C, we need simply make the changes

$$H \to C \qquad\qquad 1 \to 3 \qquad\qquad 2 \to 4$$

in our subscripts, so that

$$Q_C = W_{34} = NkT_C \ln\frac{V_4}{V_3} = -NkT_C \ln\frac{V_3}{V_4} \tag{13.47}$$

Since $V_3 > V_4$, Q_C and W_{34} are, as expected, negative: during the isothermal compression, we do work on the system, which sheds heat into the cold reservoir.

For any adiabatic process, $Q = 0$. To carry out the integration

$$W_{23} = \int_{V_2}^{V_3} p\, dV$$

for the work done during the adiabatic expansion, we need to know p as a function of V, and eq. (13.45) gives us this: since pV^γ is constant for an adiabatic process, throughout the expansion it will equal its initial value $p_2 V_2^\gamma$, so that

$$pV^\gamma = p_2 V_2^\gamma$$
$$p = \frac{p_2 V_2^\gamma}{V^\gamma}$$

and

$$\begin{aligned} W_{23} &= \int_{V_2}^{V_3} \frac{p_2 V_2^\gamma}{V^\gamma}\, dV \\ &= p_2 V_2^\gamma \int_{V_2}^{V_3} V^{-\gamma}\, dV \\ &= p_2 V_2^\gamma \left[\frac{1}{1-\gamma} V^{1-\gamma} \right]_{V_2}^{V_3} \\ &= p_2 V_2^\gamma \frac{1}{1-\gamma} \left(V_3^{1-\gamma} - V_2^{1-\gamma} \right) \\ &= \frac{1}{1-\gamma} p_2 V_2^\gamma V_3^{1-\gamma} - \frac{1}{1-\gamma} p_2 V_2 \end{aligned} \tag{13.48}$$

This is ugly but can be simplified considerably. Applying $pV^\gamma = p_2 V_2^\gamma$ at the endpoint of the expansion (vertex 3) gives

$$p_3 V_3^\gamma = p_2 V_2^\gamma \tag{13.49a}$$

And applying $pV = NkT$ at the start and end of the expansion (vertices 2 and 3), where $T = T_H$ and $T = T_C$, respectively, we also have

$$p_2 V_2 = NkT_H \qquad\qquad p_3 V_3 = NkT_C \tag{13.49b}$$

Using eqq. (13.49), we can rewrite eq. (13.48) as

$$W_{23} = \frac{1}{1-\gamma} p_3 V_3^\gamma V_3^{1-\gamma} - \frac{1}{1-\gamma} p_2 V_2$$

$$
\begin{aligned}
&= \frac{1}{1-\gamma}\, p_3 V_3 - \frac{1}{1-\gamma}\, p_2 V_2 \\
&= \frac{1}{1-\gamma}\, NkT_C - \frac{1}{1-\gamma}\, NkT_H \\
&= \frac{1}{1-\gamma}\, Nk(T_C - T_H) \\
&= \frac{1}{\gamma-1}\, Nk(T_H - T_C)
\end{aligned}
\tag{13.50}
$$

Since $\gamma > 1$ and $T_H > T_C$, the work done by the system during the adiabatic expansion is, as expected, positive.

Since during the adiabatic compression we go from T_C to T_H rather than from T_H to T_C, the work done during the adiabatic compression differs only in sign from that done during the adiabatic expansion:

$$
W_{41} = \frac{1}{\gamma-1}\, Nk(T_C - T_H) = -\frac{1}{\gamma-1}\, Nk(T_H - T_C)
\tag{13.51}
$$

This had to be so because of the first law: since ΔU vanishes for each of the isothermal legs and must also vanish for the cycle overall,

$$
\begin{aligned}
0 &= \Delta U_{cycle} \\
&= \Delta U_{12} + \Delta U_{23} + \Delta U_{34} + \Delta U_{41} \\
&= 0 + \Delta U_{23} + 0 + \Delta U_{41}
\end{aligned}
$$

which, if we apply $\Delta U = Q - W$ with $Q = 0$ to the adiabatic legs 23 and 41, yields

$$
\begin{aligned}
0 &= (Q_{23} - W_{23}) + (Q_{41} - W_{41}) \\
&= (0 - W_{23}) + (0 - W_{41}) \\
&= -(W_{23} + W_{41})
\end{aligned}
$$

and thus $W_{41} = -W_{23}$.

Pulling together what we have so far, the net work done by the system during the cycle is, by eqq. (13.46), (13.47), (13.50), and (13.51),

$$
\begin{aligned}
W_{net} &= W_{12} + W_{23} + W_{34} + W_{41} \\
&= NkT_H \ln\frac{V_2}{V_1} + \frac{1}{\gamma-1}\, Nk(T_H - T_C) \\
&\qquad - NkT_C \ln\frac{V_3}{V_4} - \frac{1}{\gamma-1}\, Nk(T_H - T_C) \\
&= NkT_H \ln\frac{V_2}{V_1} - NkT_C \ln\frac{V_3}{V_4}
\end{aligned}
\tag{13.52a}
$$

while heat

$$Q_H = NkT_H \ln \frac{V_2}{V_1} \tag{13.52b}$$

is taken from the hot reservoir and heat

$$Q_C = -NkT_C \ln \frac{V_3}{V_4} \tag{13.52c}$$

is, by virtue of being negative, dumped into the cold reservoir.

These results can, however, be simplified still further. According to eq. (13.44), $TV^{\gamma-1}$ is constant throughout an adiabatic process, and this gives us two relations among the volumes V_1, V_2, V_3, and V_4: for the adiabatic expansion, which takes us from temperature T_H at vertex 2 to temperature T_C at vertex 3,

$$T_H V_2^{\gamma-1} = T_C V_3^{\gamma-1}$$

and for the adiabatic compression, which takes us from temperature T_C at vertex 4 back to temperature T_H at vertex 1,

$$T_C V_4^{\gamma-1} = T_H V_1^{\gamma-1}$$

We can obtain results for the ratios of volumes that occur in eqq. (13.52) if we divide both sides of these relations by $T_C V_2^{\gamma-1}$ and $T_H V_4^{\gamma-1}$, respectively:

$$\frac{T_H V_2^{\gamma-1}}{T_C V_2^{\gamma-1}} = \frac{T_C V_3^{\gamma-1}}{T_C V_2^{\gamma-1}} \qquad \frac{T_C V_4^{\gamma-1}}{T_H V_4^{\gamma-1}} = \frac{T_H V_1^{\gamma-1}}{T_H V_4^{\gamma-1}}$$

$$\frac{T_H}{T_C} = \left(\frac{V_3}{V_2}\right)^{\gamma-1} \qquad \frac{T_C}{T_H} = \left(\frac{V_1}{V_4}\right)^{\gamma-1}$$

and hence

$$\frac{V_3}{V_2} = \left(\frac{T_H}{T_C}\right)^{1/(\gamma-1)} \qquad \frac{V_4}{V_1} = \left(\frac{T_H}{T_C}\right)^{1/(\gamma-1)} \tag{13.53}$$

Eqq. (13.53) are telling us that, as we had surmised earlier, only two of the four volumes V_1, V_2, V_3, and V_4 can be chosen arbitrarily, and the values of the other two are then determined. In particular, combining eqq. (13.53) gives

$$\frac{V_2}{V_1} = \frac{V_3}{V_4} \tag{13.54}$$

which allows us to simplify eqq. (13.52) to

$$\begin{aligned} W_{net} &= NkT_H \ln \frac{V_2}{V_1} - NkT_C \ln \frac{V_3}{V_4} \\ &= NkT_H \ln \frac{V_2}{V_1} - NkT_C \ln \frac{V_2}{V_1} \end{aligned}$$

$$= Nk(T_H - T_C) \ln \frac{V_2}{V_1} \tag{13.55a}$$

$$Q_H = NkT_H \ln \frac{V_2}{V_1} \tag{13.55b}$$

$$Q_C = -NkT_C \ln \frac{V_3}{V_4}$$

$$= -NkT_C \ln \frac{V_2}{V_1} \tag{13.55c}$$

This is probably a good point to summarize and take stock of what we have found so far. First, note that by eqq. (13.55),

$$W_{net} = Q_H + Q_C \tag{13.56}$$

This is what we would expect from the first law: for the cycle overall,

$$\begin{aligned} \Delta U_{cycle} &= Q_{net} - W_{net} \\ 0 &= (Q_H + Q_C) - W_{net} \\ W_{net} &= Q_H + Q_C \quad \checkmark \end{aligned}$$

Next, note that only during the isothermal expansion does the system absorb heat: on the two adiabatic legs, $Q = 0$, and during the isothermal compression the system sheds heat. The heat Q_H absorbed during the isothermal expansion therefore constitutes the entirety of the heat used to run a Carnot engine. During each iteration of the cycle, some of this heat energy is converted into the net work W_{net} done by the engine and the rest, Q_C, is dumped into the cold reservoir as waste heat. The more efficient a heat engine, the more work we get out of it for the heat energy used to run it; quantitatively, we would therefore define the efficiency η of a heat engine — any heat engine, not just a Carnot engine — as

$$\eta = \frac{W_{net}}{Q_H} \tag{13.57}$$

where Q_H would, more generally, be defined as the total heat taken in by the engine. If all of the heat taken in were converted to work, the efficiency would be unity, that is, 100%. By eqq. (13.55a) and (13.55b), the efficiency of a Carnot engine is

$$\eta_{Carnot} = \frac{Nk(T_H - T_C) \ln \frac{V_2}{V_1}}{NkT_H \ln \frac{V_2}{V_1}}$$

$$= 1 - \frac{T_C}{T_H} \tag{13.58}$$

which is less than 100% whenever the cold reservoir T_C is above absolute zero.

To maximize the efficiency of a Carnot engine, you need to minimize the subtracted term T_C/T_H by making T_C as small as possible and T_H as large as possible, that is, by using the coldest cold reservoir and the hottest hot reservoir available. While the efficiencies of other sorts of heat engines are not given by eq. (13.58), they are generally also more efficient for more extreme temperatures, and this is why power plants use very hot steam, produced by burning fossil or nuclear fuel, to drive the turbines that generate power.

Finally, taking the ratio of eqq. (13.55b) and (13.55c) yields

$$\begin{aligned} \frac{Q_H}{Q_C} &= \frac{NkT_H \ln \frac{V_2}{V_1}}{-NkT_C \ln \frac{V_2}{V_1}} \\ &= -\frac{T_H}{T_C} \end{aligned} \tag{13.59}$$

This humble-looking relation actually turns out to be of great fundamental importance, principally for reasons we will get to in §13.8, but also because it gives us a rigorous way to measure temperature: in principle, we can relate the temperatures of two bodies by running a Carnot cycle between them and measuring Q_H and Q_C.[37]

13.7.2 Air Conditioning & Refrigeration

A refrigerator is just a heat engine run backward: instead of putting heat into the engine and getting work out of it, you put work into running the refrigerator and it removes heat for you. In particular, if we go around the Carnot cycle of fig. (13.3) on p.705 counterclockwise rather than clockwise, the values of p, V, and T at each vertex will of course be the same, but the signs on the values of ΔU, Q, and W will be reversed, so that eqq. (13.55) become

$$W_{net} = -Nk(T_H - T_C) \ln \frac{V_2}{V_1} < 0 \tag{13.60a}$$

$$Q_H = -NkT_H \ln \frac{V_2}{V_1} < 0 \tag{13.60b}$$

$$Q_C = NkT_C \ln \frac{V_2}{V_1} > 0 \tag{13.60c}$$

[37] The heat energies Q_H and Q_C that the cycle draws from the hotter body and dumps into the colder body would, however, have to be infinitesimal when the temperatures of finite bodies are being measured; otherwise the temperatures of these bodies, which constitute the hot and cold reservoirs, would not remain constant, and the isothermal legs of the Carnot cycle would no longer be isothermal.

Since W_{net} is the work done *by* the system, a negative W_{net} means that we have to do work on the system to run the cycle. And instead of heat being absorbed from the hot reservoir and dumped in the cold reservoir, our positive Q_C and negative Q_H indicate that the refrigerator absorbs heat from the cold reservoir and dumps heat into the hot reservoir. In terms of the big white box sitting in your kitchen, the cold reservoir is the interior of the fridge and the hot reservoir is the room, that is, the kitchen. The cycle of your fridge is not a Carnot cycle, but the basic scheme is the same: The work required to run the cycle is supplied by an electric motor that compresses a gas into coiled tubes on the exterior of the rear of the fridge. As a result of this compression, the temperature of the gas rises. A fan blows air over these coils to cool the compressed gas, and when its temperature has dropped enough, a valve opens and the gas adiabatically expands into another set of coiled tubes on the interior of the fridge. As a result of this expansion, the temperature of the gas drops. Another fan cools the air in the interior of the fridge by circulating it over these cooled coils. The gas is then pumped from these interior coils back into the exterior coils by the electric compressor and the cycle repeats. You have probably heard the the compressor kick in and rumble away when you have walked by the fridge during that part of the cycle, and you have probably also heard the interior fan when you have opened the fridge during the cooling stage. The fan that blows air over the exterior coils is usually quieter, but you can hear it if you get close and listen carefully. You'd have to be really, really bored to bother doing that, but in principle you could do it.

At any rate, for a Carnot refrigerator we can rewrite eq. (13.56),

$$W_{net} = Q_H + Q_C$$

in the form

$$\begin{aligned} Q_H &= -Q_C + W_{net} \\ &= -Q_C - W_{us} \\ &= -(Q_C + W_{us}) \end{aligned} \tag{13.61}$$

where $W_{us} = -W_{net}$ is the work we must do to run the cycle. Eq. (13.61) is telling is that the cycle takes the heat energy it has withdrawn from the cold reservoir and dumps it, together with the energy we have put into running the cycle, into the hot reservoir. This also holds true for real refrigerators, even though their cycle is not a Carnot cycle. Air conditioners, for example, remove heat from your home and dump that heat, together with the work done courtesy of the power company to run the air conditioner, into the great outdoors. This is in fact why air conditioners always have to be partly inside and partly outside: an air conditioner that was entirely inside a room would

just be dumping the heat back into the room — which, since this dumped heat energy includes the work needed to run the cycle, would have the net effect of heating the room up instead of cooling it off.[38]

The more efficient a refrigerator, the more heat it withdraws from the cold reservoir for the work done to run it. We would therefore logically define the efficiency η of a refrigerator — any refrigerator, not just a Carnot refrigerator — as

$$\eta = \frac{Q_C}{W_{us}} \tag{13.62}$$

By eqq. (13.60a), (13.60c), and $W_{us} = -W_{net}$, the efficiency of a Carnot refrigerator is

$$\begin{aligned} \eta_{Carnot} &= \frac{NkT_C \ln \frac{V_2}{V_1}}{Nk(T_H - T_C) \ln \frac{V_2}{V_1}} \\ &= \frac{T_C}{T_H - T_C} \end{aligned} \tag{13.63}$$

A few things to note about the efficiency of a refrigerator:

- It is not only possible but common for the efficiency of a refrigerator to exceed 100%: an $\eta > 1$ in eq. (13.63) simply means that $Q_C > W_{us}$, that is, that every Joule of work we do to run the refrigerator removes more than one Joule of heat from the cold reservoir.[39]

- The lower the temperature T_C of the cold reservoir, the less efficient the refrigerator. In fact, as $T_C \to 0$, $\eta \to 0$: even in principle you can

[38] Actually, you can have an air conditioner that is entirely inside the room if you can manage to store the heat somewhere. We'll look at this trick in problem #19.

[39] Air conditioners and refrigerators are labeled with an "energy efficiency ratio" (EER) or "seasonal energy efficiency ratio" (SEER). The logical definition of efficiency as the ratio of heat energy removed to work done to run the cycle is a dimensionless ratio, that is, a pure number, and anyone in his or her right mind would of course calculate this pure number using the same energy units for the heat energy in the numerator and the work in the denominator. Engineers have, however, for some perverse reasons that we are probably all better off not knowing, decided to define the EER as the ratio of the BTUs of heat energy removed from the cold reservoir to the Watt-hours of energy required to run the unit between two wistfully chosen temperatures that are supposed to apply to everyone, even though the actual operating temperatures obviously vary widely with geographical location. (A BTU (British thermal unit) is a unit of energy: 1 BTU = 1055 J.) But while pretty useless for determining the efficiency you will get from a given air conditioner, EERs and SEERs are a valid indication of the *relative* efficiency of different air conditioners. Glancing at the plate on the side of the air conditioner humming away as we write this, for example, we see that the unit claims to remove heat from the room at 10,200 BTU/hr, to draw 940 W, and to have an EER of 10.8, and in fact $10,200/940 = 10.85$. Fat chance that this air conditioner is actually operating with that efficiency for us, right here, on this particular day, but we can be confident that it is operating more efficiently for us than would a unit with an EER of 10 and less efficiently than would a unit with an EER of 12.

never cool anything down to absolute zero, because it would end up taking an infinite amount of energy to get there. Big-deal refrigerators in cryogenics labs can get down to small fractions of a degree Kelvin, but it's impossible to get all the way down to 0 K.

- The closer the temperatures T_H and T_C between which the refrigerator is operating, the more efficient the refrigerator. In fact, if $T_H = T_C$, the efficiency of a Carnot refrigerator is infinite.

13.8 Reversibility, Entropy, & the Second Law

A reversible process is any process that can, at least in principle, be reversed to return both the system and its environment to their original states. Note that it is not enough that the system be returned to its original state: if after the process has been reversed there remain changes to the system's environment, then the process is not considered reversible. For example, we might allow a gas in a cylinder fitted with a freely movable piston to undergo a slow isothermal expansion. During this expansion, the volume of the gas increases, its pressure decreases, it does work, and it absorbs heat energy from its environment. If we reverse this process and do a slow isothermal compression, the gas will return to its original volume and pressure, we will do on the gas the same amount of work that it did during the expansion, and the gas will shed back into its environment the same amount of heat that it absorbed from it during the expansion. As a result, both the gas and its environment have been returned to their original state, and the isothermal expansion is a reversible process. If, on the other hand, the piston is not freely movable but experiences friction, we will have to do additional work to overcome friction during the expansion and compression, work that will be converted by the friction into heat. Even if the gas is returned to its original state, this additional heat due to friction will have been dumped into the environment, which is therefore not in its original state. When there is friction, the expansion is irreversible.

Two other classic examples of irreversible processes: An adiabatic process may be reversible or irreversible, depending on how it is carried out. If an adiabatic expansion or compression of a gas in a cylinder fitted with a freely movable piston is carried out too quickly, it will be irreversible: the effects of the change will not have time to propagate and distribute themselves uniformly throughout the volume of the gas, with the result that the gas is left in a nonequilibrium state in which its temperature and pressure are not even well defined. Such a process is not quasistatic and could not be

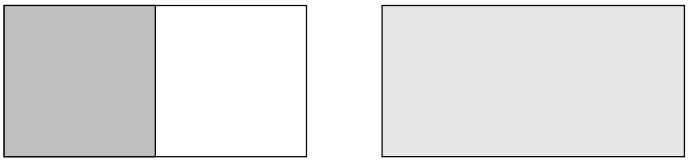


Figure 13.6: *Freedom!*

represented by a curve in the pV plane. And if, as shown in fig. (13.6), a membrane confining a gas to one side of a box is suddenly removed, the effect is similarly irreversible: the gas molecules will spontaneously distribute themselves throughout the entire volume of the box, and those chickens just ain't goin' back in the hen house. You could, of course, compress the gas back into its original volume, but this would not return the system to its original state because of the energy added to the system by virtue of the work you have done compressing the gas.

To be reversible, a process must be continuous, quasistatic, and nondissipative — that is, without loss of energy to friction — , so that we can reverse the process by reversing the succession of infinitesimally differing equilibrium states through which it was carried out. All of the processes with which we have dealt in previous sections satisfy these criteria. In particular, the isothermal and adiabatic expansions and compressions of the Carnot cycle, and thus the Carnot cycle itself, are reversible. In fact, any reversible cycle can be regarded as a superposition of Carnot cycles: If we think in terms of closed loops in the pV plane, a loop of any shape can be regarded as a superposition of infinitesimal rectangular cycles, as crudely shown for rectangular cycles that are decidedly not infinitesimal in fig. (13.7). In the limit that the rectangles become infinitesimally small, they will exactly reproduce the blue cycle, and going around the blue cycle will be equivalent to going around each of these myriad infinitesimal rectangles: if, for example, we go

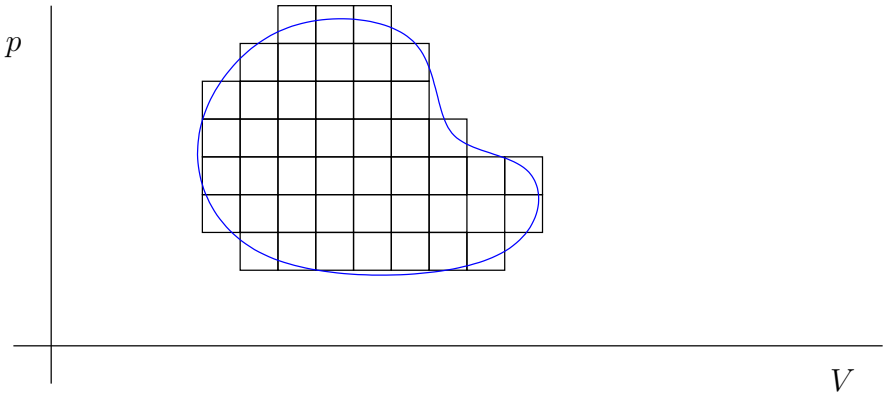


Figure 13.7: Fitting Squarish Pegs into Roundish Holes

around counterclockwise, then the left side of each rectangle will be traversed top-to-bottom and each right side bottom-to-top, with the result that on the interior, where the sides of the rectangular cycles overlap, their contributions will cancel. The only surviving contributions will be those from the exterior sides of the rectangular cycles, which together reproduce the blue cycle. Although the Carnot cycle is a much funkier shape than a rectangle,[40] any closed loop in the pV plane can be similarly reproduced by a superposition of Carnot cycles.

Which of course raises the question of why anyone would want to reproduce other cycles as superpositions of Carnot cycles. And the answer, believe it or not, would be relation (13.59):

$$\frac{Q_H}{Q_C} = -\frac{T_H}{T_C}$$

If we write this relation in the form

$$\frac{Q_H}{T_H} + \frac{Q_C}{T_C} = 0$$

and note that $Q = 0$ for the two adiabatic legs, then we see that for a Carnot cycle

$$\oint_{\text{cycle}} \frac{dQ}{T} = 0 \tag{13.64}$$

Since any reversible cycle can be regarded as a superposition of Carnot cycles, eq. (13.64) holds for any reversible cycle. This in turn means that the integral

$$\int_A^B \frac{dQ}{T} \tag{13.65}$$

over any reversible process depends only on the initial and final states A and B of that process and not on the process itself: if P_1 and P_2 are two reversible paths from A to B, then P_1 and the reverse of P_2 — which we will denote by $\widetilde{P}_2$ — together form a closed loop, so that by eq. (13.64)

$$0 = \oint_{P_1\widetilde{P}_2} \frac{dQ}{T} = \int_{P_1 A}^{B} \frac{dQ}{T} + \int_{\widetilde{P}_2 B}^{A} \frac{dQ}{T} = \int_{P_1 A}^{B} \frac{dQ}{T} - \int_{P_2 A}^{B} \frac{dQ}{T}$$

and thus

$$\int_{P_1 A}^{B} \frac{dQ}{T} = \int_{P_2 A}^{B} \frac{dQ}{T}$$

[40] This is, of course, why we drew fig. (13.7) with rectangles; we have neither the time nor the patience to draw it with Carnot cycles. And even if we did, it would have just ended up looking like a Jackson Pollock painting.

Since this holds for any reversible paths P_1 and P_2 from A to B, the integral of eq. (13.65) must be independent of path and depend only on the initial state A and final state B. This integral therefore constitutes a state variable, which we will denote by S:

$$\Delta S_{AB} = \int_A^B \frac{dQ}{T} \tag{13.66}$$

This new animal S is called entropy.

Using the definition (13.66) of entropy in eq. (13.64), we see that there is no change in entropy around any reversible cycle:

$$\oint dS = 0 \tag{13.67}$$

To see what further properties the entropy S has, consider the case of spontaneous heat flow, which is always from hotter to colder bodies. For the simple case of two bodies, a hotter body at temperature T_H and a colder body at temperature T_C, the heat dQ_C absorbed by the colder body is the heat shed by the hotter body, so that the dQ of the hotter body is $dQ_H = -dQ_C < 0$. The change in entropy is therefore positive for the colder body and negative for the hotter body:

$$dS_C = \frac{dQ_C}{T_C} > 0 \qquad\qquad dS_H = \frac{dQ_H}{T_H} = -\frac{dQ_C}{T_H} < 0$$

Because $T_H > T_C$, the overall change in entropy is, however, positive:

$$dS_{net} = dS_C + dS_H = \frac{dQ_C}{T_C} - \frac{dQ_C}{T_H} = dQ_C\left(\frac{1}{T_C} - \frac{1}{T_H}\right) > 0 \tag{13.68}$$

If we were recklessly adventurous, we might extrapolate and hypothesize that there is no change in entropy during reversible processes and that during spontaneous, irreversible processes entropy always increases and never decreases. But, hey, you only go around once, so what the ****, let's go for it. Our task is now to prove this.

Doing so will require invoking the second law of thermodynamics, and since you can't very well go around invoking laws you've never heard of, we suppose this would be a good time to say a few words about it. The second law can be — and historically was — stated in a variety of ways, including, to paraphrase loosely,[41]

[41] Our enumeration of these statements by letter and number is only to facilitate references later in the text and is *not* conventional.

1. (a) Heat flows spontaneously only from hotter to colder bodies.[42]
   (b) There is no perfect refrigerator. That is, there is no thermodynamic cycle that can transfer heat from a colder to a hotter body without work.

2. (a) There is no thermodynamic perpetual-motion machine.
   (b) There is no perfect heat engine. That is, there is no thermodynamic cycle that can convert environmental heat entirely into work.

3. Total entropy never decreases: during reversible processes it doesn't change; during spontaneous, irreversible processes it always increases.

We will assume statement (1a) to be true. With that as our starting point, we will first establish statements (1b) and (2b): Imagine a Carnot heat engine and a Carnot refrigerator operating between the same two temperatures T_H and T_C, with the refrigerator driven by the output of the engine. The refrigerator would exactly undo the effects of the engine: as much heat energy as the engine dumped into the cold reservoir, the refrigerator would remove, and the refrigerator would dump as much heat into the hot reservoir as the engine had removed. Now suppose that you replace the Carnot engine with a more efficient heat engine we will call Fred: Fred will convert a greater proportion of the heat drawn from the hot reservoir into work, so that for the same heat drawn from the hot reservoir, Fred would produce more work to drive the Carnot refrigerator and dump less waste heat into the cold reservoir. The Carnot refrigerator would then remove more heat from the cold reservoir than Fred dumped into the cold reservoir, and dump more heat into the hot reservoir than Fred removed from the hot reservoir. The net effect of the tandem operation of Fred and the Carnot refrigerator would be the transfer of heat from the cold to the hot reservoir, violating statement (1a). There is therefore no heat engine more efficient than a Carnot engine. And since a refrigerator is just a heat engine run backward, there is therefore also no refrigerator more efficient than a Carnot refrigerator. Since it takes work to run a Carnot refrigerator, statement (1b) is established. And since not all of the heat absorbed by a Carnot heat engine is converted into work, statement (2b) is also established.

This is in fact the reason for the great emphasis given to the Carnot cycle in thermodynamics: it is the most efficient possible cycle. Note also that this limit on efficiency is a limit in principle, not a practical consequence of the performance of real engines falling short of the ideal because of friction or less than perfectly machined parts or lack of some engineering technology: you can no more design a heat engine more efficient than a Carnot engine

[42] *Spontaneously* meaning of its own accord, without requiring work.

than you can trisect an angle with a straightedge and compass, and even Santa Claus and the Easter Bunny can't do that.

Anyway, there being no perfect heat engine, a thermodynamic perpetual-motion machine is out of the question: since some of the heat drawn from the hot reservoir is necessarily dumped as waste heat into the cold reservoir, any two real reservoirs will eventually even out at a common temperature, at which point, by eq. (13.58), even the efficiency

$$\eta = 1 - \frac{T_C}{T_H}$$

of a Carnot engine will drop to zero and the perpetual-motion machine will cease moving perpetually. Statement (2a) is thus established.[43]

Which brings us back to the issue of entropy. To show that there is no change in the total entropy of the system and its environment during a reversible process, recall that any reversible cycle can be reproduced by a superposition of Carnot cycles. Each leg of a reversible cycle — that is, a reversible process — is therefore equivalent to a superposition of the various legs of the Carnot cycle. On the adiabatic legs of the Carnot cycle, $dQ = 0$, so that $dS = dQ/T = 0$. And on the isothermal legs, the heat dQ absorbed by the system is drawn from the reservoir, so that the heat absorbed by the reservoir is $dQ_r = -dQ$. Since on an isothermal leg the system and the reservoir are at the same temperature T — either T_H or T_C — , the net change in entropy along any segment of an isothermal leg is

$$dS = \frac{dQ}{T} + \frac{dQ_r}{T} = \frac{dQ}{T} - \frac{dQ}{T} = 0$$

For a reversible process there is therefore no change in total entropy.

To demonstrate that irreversible processes increase the total entropy takes a bit more work. Consider a heat-engine cycle that operates between a high temperature T_H and a low temperature T_C. If any part of this cycle is irreversible, its efficiency $\eta = W_{net}/Q_H$ must be less than the Carnot efficiency $1 - T_C/T_H$ of eq. (13.58):[44]

$$\frac{W_{net}}{Q_H} < 1 - \frac{T_C}{T_H} \tag{13.69}$$

Using $W_{net} = Q_H + Q_C$ from eq. (13.56) in eq. (13.69), we have

$$\frac{Q_H + Q_C}{Q_H} < 1 - \frac{T_C}{T_H}$$

[43] Although you could still, as pointed out by our former student Trey Kollmer, have a perpetual-motion machine that "works for a while."

[44] Our irreversible cycle may, of course, take heat in at other than its highest temperature T_H and shed heat at other than its lowest temperature T_C, but to the extent that it does so it will be even less efficient. We may therefore assume that all of the heat Q_H absorbed by the engine is absorbed at T_H and that all of the heat Q_C shed by it is shed at T_C.

$$\begin{aligned} 1 + \frac{Q_C}{Q_H} &< 1 - \frac{T_C}{T_H} \\ \frac{Q_C}{Q_H} &< -\frac{T_C}{T_H} \\ \frac{Q_C}{T_C} &< -\frac{Q_H}{T_H} \\ \frac{Q_C}{T_C} + \frac{Q_H}{T_H} &< 0 \end{aligned}$$

or, expressed as an integral around the cycle,

$$\oint \frac{dQ}{T} < 0 \tag{13.70}$$

Eq. (13.70) might tempt you to conclude that the entropy change around an irreversible cycle is negative, that is, that entropy decreases. This would, however, be forgetting that the integral of dQ/T on the right-hand side of eq. (13.66) gives the change in entropy *only when the path of integration is reversible*. To see the effect of an irreversible process on entropy, we need to pair an irreversible process P_i that takes us from state A to state B with a reversible process P_r that takes us between those same two states, so that we can use eq. (13.66) to calculate the entropy change along the reversible path: since entropy is a state variable, the entropy change along the irreversible path will be the same as the change along the reversible path. In particular, if P_i and P_r both take us from A to B, and if we denote the reverse of the path P_r by $\widetilde{P}_r$, together P_i and $\widetilde{P}_r$ form a closed loop to which eq. (13.70) will apply:

$$0 > \oint_{\widetilde{P}_r P_i} \frac{dQ}{T} = \int_{\widetilde{P}_r}^{B} \frac{dQ}{T} + \int_{P_i}^{B} \frac{dQ}{T} = -\int_{P_r}^{B} \frac{dQ}{T} + \int_{P_i}^{B} \frac{dQ}{T}$$

and hence, if we apply the definition (13.66) of entropy to the integral over the reversible path P_r,

$$0 > -\Delta S_{AB} + \int_{P_i}^{B} \frac{dQ}{T}$$

Thus

$$\Delta S_{AB} > \int_{P_i}^{B} \frac{dQ}{T}$$

For a thermally isolated system, or equivalently for a system and its environment together, all of the contributions to dQ on the right-hand side are from internal exchanges — the heat shed by one body is absorbed by others

— so that the net dQ vanishes and we have [45]

$$\Delta S > 0 \tag{13.71}$$

Entropy is therefore always increased by irreversible processes.

The grandest isolated system to which eq. (13.71) applies is the universe itself, the entropy of which can never decrease and is in fact always increasing because of irreversible processes like the flow of heat from hotter to colder bodies. If, as may be the case — astrophysical observations are not yet accurate enough to know for sure [46] — , the universe continues to expand forever without collapsing back in on itself, this continual increase in the entropy will ultimately lead to its heat death as it evens out into an ever thinner and more uniform gray mush ever nearer absolute zero. Without temperature differences that would enable heat engines to do work and make things happen, life will no longer be possible, nor much of anything, for that matter. Schopenhauer would be delighted.[47]

In the meantime, it is important to note there is no reason why the entropy of a system or body cannot decrease. When, for example, heat flows from a hotter to a colder body, the entropy of the hotter body decreases because its dQ/T is negative. Similarly when a Carnot refrigerator removes heat from the cold reservoir, the reservoir's dQ/T and change in entropy are negative. What cannot decrease is the *total* entropy: for reversible processes and cycles there is no overall change in entropy, and for irreversible processes and cycles there is an overall increase. In the case of the Carnot refrigerator, the decrease in entropy when heat is drawn from the cold reservoir is canceled by an equal increase in entropy when this heat and the work done to run the cycle are dumped into the hot reservoir. And in the case of heat flow from a hotter to a colder body, the increase in the entropy of the colder body is, as we have seen, greater than the decrease in the entropy of the hotter body.

[45] You could quite legitimately object that the dQ is divided by T and that the various parts of a system and its environment need not be at the same temperature, but in the absence of work done by an external agent heat would flow only from hotter to colder bodies, which we saw in eq. (13.68) yields a net increase in entropy.

[46] Reading about stuff like black holes, quasars, collapsing universes, and the Big Bang is fascinating; if you haven't already, you will doubtless read a great deal of astrophysics in magazines, newspapers, and popular books. When you do, you should, however, bear in mind that astrophysicists, on a scale of trustworthiness, come just after used-car salesmen: devoted to a profession by its very nature far removed from reality, they are given to flights of fancy, and will present the most arrant and fantastic speculation as though it were established by wide scientific consensus. But how much credence should we give to people who can't even find 90% of the mass in the universe?

[47] We suppose you could start an environmental movement to conserve entropy, setting up tables at supermarkets and trying to educate people about the dangers of thermodynamically irreversible processes, lobbying legislators for regulations limiting entropy increases, and making frothy emotional appeals to people to think of their children and grandchildren.

As we saw from statement (3) on p.721, entropy provides us with another way of expressing the second law. And as we saw in our arguments justifying that expression of the second law, entropy also provides us with an alternative definition of reversibility: a reversible process is one that does not change the total entropy; an irreversible process is one that increases the total entropy. Historically, the definition of entropy and its relevance to the second law and reversibility were the result of a long, tortuous, painful effort to understand observed thermodynamic behavior, perhaps not dissimilar to what you have experienced reading this section. The more modern perspective starts from entropy and states the second law and the concept of reversibility in terms of it. Which makes it all the more poignant that we have yet to say anything about what entropy actually is — a situation we will attempt to remedy in the next section.

13.8.1 Entropy in Statistical Mechanics

Each macroscopic state of a system corresponds to numerous possible microscopic states: an ideal gas occupying a box of volume V at pressure p and temperature T is in a unique macroscopic state, but there are myriad microscopic states, each with a different distribution of the locations and velocities of the N molecules in the sample, that could give rise to that macroscopic state. Statistical mechanics assumes that the possible microscopic states of a system are all equally probable. If in addition we assume that the entropy S of a macroscopic state of a system Σ is a function of the number $\mathcal{N}$ of microscopic states that correspond to that macroscopic state, then we can derive a result for S by thinking of Σ as two subsystems, Σ_1 and Σ_2. If Σ is a gas in a box, for example, Σ_1 might be the gas in some subregion of the box and Σ_2 the rest of the gas. Now, entropy should, as we saw in the preceding section, be additive: the entropy of Σ should be the sum of the entropies of Σ_1 and Σ_2. The multiplicity of microscopic states is, however, multiplicative: if $\mathcal{N}_1$ microscopic states give rise to Σ_1 and $\mathcal{N}_2$ to Σ_2, then there are $\mathcal{N} = \mathcal{N}_1\mathcal{N}_2$ microscopic states that could give rise to Σ. If the entropies of Σ, Σ_1, and Σ_2 are functions of $\mathcal{N}$, $\mathcal{N}_1$, and $\mathcal{N}_2$, respectively, then

$$S_\Sigma = S_{\Sigma_1} + S_{\Sigma_2}$$

will therefore take the form

$$S(\mathcal{N}_1\mathcal{N}_2) = S(\mathcal{N}_1) + S(\mathcal{N}_2) \tag{13.72}$$

Even without solving eq. (13.72), we can begin to see how S behaves: for $\mathcal{N}_1 = \mathcal{N}_2 = \mathcal{N}$,

$$S(\mathcal{N}^2) = S(\mathcal{N}) + S(\mathcal{N}) = 2S(\mathcal{N})$$

which suggests a logarithm. In particular, $S(1^2) = 2S(1)$ yields

$$S(1) = 0 \tag{13.73}$$

To generate a differential equation that we can actually solve for S, we can look at two infinitesimally differing values of $\mathcal{N}$: with $\mathcal{N}_1 = \mathcal{N}$ and $\mathcal{N}_2 = 1 + \epsilon$, eq. (13.72) becomes

$$\begin{aligned} S\Big(\mathcal{N}(1+\epsilon)\Big) &= S(\mathcal{N}) + S(1+\epsilon) \\ S(\mathcal{N} + \epsilon\mathcal{N}) &= S(\mathcal{N}) + S(1+\epsilon) \end{aligned}$$

A Taylor expansion of this to first order in the infinitesimal ϵ yields

$$S(\mathcal{N}) + \epsilon\mathcal{N}S'(\mathcal{N}) = S(\mathcal{N}) + S(1) + \epsilon S'(1)$$

where S' is the derivative of S. If we cancel the $S(\mathcal{N})$ terms on the left and right sides and use $S(1) = 0$ from eq. (13.73), we obtain

$$\begin{aligned} \epsilon\mathcal{N}S'(\mathcal{N}) &= \epsilon S'(1) \\ S'(\mathcal{N}) &= \frac{S'(1)}{\mathcal{N}} \\ \frac{dS}{d\mathcal{N}} &= \frac{S'(1)}{\mathcal{N}} \\ \int_0^S dS &= \int_1^{\mathcal{N}} S'(1)\,\frac{d\mathcal{N}}{\mathcal{N}} \end{aligned}$$

where in the lower limits we have noted that the value of S corresponding to $\mathcal{N} = 1$ is $S(1) = 0$. Although we don't know the value of $S'(1)$, we do know that it is just a constant that can be pulled out of the integration, so that

$$\begin{aligned} S &= S'(1)\Big[\ln\mathcal{N}\Big]_1^{\mathcal{N}} \\ &= S'(1)\ln\mathcal{N} \\ &= (\text{const})\ln\mathcal{N} \end{aligned} \tag{13.74}$$

The one remaining issue is the value of the constant in front of the $\ln\mathcal{N}$, and we will see in the examples of the next section that this statistical mechanical result for S will match up with our previous thermodynamic result for S if we take this constant to be the Boltzmann constant k:

$$S = k\ln\mathcal{N} \tag{13.75}$$

Eq. (13.75) gives us a new perspective on entropy, one very different from — and much more enlightening than — that of $dS = dQ/T$: we now see that

the entropy of a macroscopic state of a system grows with the number of microscopic states that give rise to it. Since each microscopic state is equally likely, the more microscopic states that give rise to a macroscopic state, the more likely that macroscopic state. Eq. (13.75) is telling us that the more likely a macroscopic state, the higher its entropy.

As a rather artificial example, consider the game of the cow and the farmer played with decks of cards of various sizes. This classic game, you will recall, is a "contest" between you and a gullible opponent: you offer to be the cow and to let your opponent be the farmer, which always sounds like a good deal to someone gullible. You then hurl the deck of cards into the air, declaring "I'm the cow and I made the mess. Now you clean it up." Your opponent has lost the game but in the process learned a valuable lesson in life.[48]

Anyway, the system in this case is the deck of cards, with the microscopic states being the various possible orderings of the cards after they have been picked up. If the deck has only two cards, there are only two possible states: one with the cards in order, the other with the cards out of order. The system is equally likely to be in either state, and, there being only one microscopic state corresponding to each of these macroscopic states, their entropies are both

$$S = k \ln 1 = 0$$

If instead the deck has three cards, there are $3! = 6$ microscopic states, for only one of which the cards are in order. Ending up in a "disordered" state is therefore five times more likely than ending up with the cards in order, and the corresponding entropies are

$$S_{ordered} = k \ln 1 = 0 \qquad\qquad S_{disordered} = k \ln 5$$

For a normal deck of 52 cards, there are $52! \approx 8 \times 10^{67}$ microscopic states, for only one of which the cards are in order. Ending up in a disordered state is therefore approximately 8×10^{67} times more likely than ending up with the cards in order, and the corresponding entropies are

$$S_{ordered} = k \ln 1 = 0 \qquad\qquad S_{disordered} \approx k \ln 8 \times 10^{67} \approx 160k$$

Of course, among the $52! - 1$ disordered states in this last case, there are many states that have very ordered subsets of the 52 cards, but most these $52! - 1$ states will be what you would consider highly disordered. The point is that as the size of a system grows, the multiplicity of microscopic states that give rise to its macroscopic states grows dramatically and becomes very lopsided, so that it becomes a virtual certainty that the system will end up the state of highest probability and thus highest entropy. This is true for a

[48] Not that we're bitter about it, mind you.

deck of 52 cards, and is astronomically more true[49] when dealing with on the order of Avogadro's number of molecules.

Suppose, for example, we have N molecules of gas in a rectangular box of volume V. If we are concerned with which half of the box each molecule is in, there are two possible states for each molecule — it's either in one half of the box or the other — , and thus 2^N possible states of the system of N molecules. There are only two states in which all the molecules are on one side of the box or the other. The probability p and entropy S of such a macroscopic state are therefore

$$p = \frac{2}{2^N} = \frac{1}{2^{N-1}} \qquad\qquad S = k \ln 2$$

Although many of the $2^N - 2$ remaining states of the gas are heavily biased toward one half of the box or the other, for the vast majority of these remaining states the gas molecules are fairly evenly distributed. The probability and entropy of such a macroscopic state are therefore on the order of

$$p = \frac{2^N - 2}{2^N} = 1 - \frac{1}{2^{N-1}}$$
$$S = k \ln(2^N - 2) \approx k \ln 2^N = Nk \ln 2$$

The molecules are thus astronomically more likely to be fairly evenly distributed throughout the box than to be found all on one side of it, and the entropy of such an even distribution is correspondingly astronomically greater.

The second law and the tendency toward higher entropy can therefore be regarded as a matter of probability: if a system starts in a state of relatively low probability and low entropy, it will tend to evolve toward states of higher probability and higher entropy, and ultimately to end up in the state of highest probability and highest entropy. Which leads us to two important observations: First, the second law, unlike the first law, is statistical in nature: it is more likely that a system will move to a state of higher probability and entropy — astronomically more likely when dealing with on the order of Avogadro's number of molecules — , but it is still possible that the system could move to a state of lower entropy. That is, the second law could be spontaneously violated; it is just that for systems of significant size such a violation is extremely unlikely.

Second, as we saw in the cow-and-the-farmer example, it is disordered states that are more numerous and therefore more likely. In fact, people often

[49] It used to be that something was either true or false, but in recent years we have learned from those who lead our societies that there are — *mirabile dictu* — degrees of truth. In fact, sometimes something can be true even though every informed person and all of the evidence in the universe say otherwise.

equate entropy with disorder and think of the second law as the tendency of systems to become more disordered. This alternative way of thinking about entropy is often useful, but should not cause you to forget that entropy is really a matter of probability and the tendency of systems evolve toward the most probable macroscopic state.

13.8.2 Some Examples & Observations

In order to match up the statistical mechanical entropy

$$S = k \ln \mathcal{N}$$

of eq. (13.75) with the thermodynamic entropy

$$\Delta S_{AB} = \int_A^B \frac{dQ}{T}$$

of eq. (13.66), consider an ideal gas with D degrees of freedom. By eq. (13.35), the internal energy of such a gas is

$$U = \tfrac{1}{2} DNkT$$

If we use the first law $dU = dQ - dW$ in the form

$$dQ = dU + dW$$

and $dS = dQ/T$ in the form

$$dQ = T\, dS$$

we therefore have

$$\begin{aligned} dQ &= dU + dW \\ T\, dS &= d\Big(\tfrac{1}{2} DNkT\Big) + dW \\ &= \tfrac{1}{2} DNk\, dT + dW \\ dS &= \frac{\frac{1}{2} DNk}{T}\, dT + \frac{dW}{T} \end{aligned} \tag{13.76}$$

In order to integrate this relation and get a result for S, we need to gerrymander the dW/T term into an integrable form, which we can accomplish by using $dW = p\, dV$ and the ideal-gas law in the form

$$p = \frac{NkT}{V}$$

Eq. (13.76) then yields

$$\begin{aligned} dS &= \frac{\frac{1}{2}DNk}{T}\,dT + \frac{p\,dV}{T} \\ &= \frac{\frac{1}{2}DNk}{T}\,dT + \frac{NkT}{V}\,\frac{dV}{T} \\ &= \tfrac{1}{2}DNk\,\frac{dT}{T} + Nk\,\frac{dV}{V} \\ \int_{S_0}^{S} dS &= \tfrac{1}{2}DNk\int_{T_0}^{T}\frac{dT}{T} + Nk\int_{V_0}^{V}\frac{dV}{V} \\ S - S_0 &= \tfrac{1}{2}DNk\ln\frac{T}{T_0} + Nk\ln\frac{V}{V_0} \end{aligned} \tag{13.77}$$

Having obtained this thermodynamic result for the entropy of an ideal gas, we are now in a position to compare it to the statistical mechanical result $S = k\ln\mathcal{N}$. If we double the volume V occupied by the gas without changing its temperature, either by an isothermal expansion or, as we did on p.718, by removing a membrane that confined the gas to one half of a box, then the change ΔS in entropy according to eq. (13.77) is

$$\begin{aligned} \Delta S &= \Delta\Big(\tfrac{1}{2}DNk\ln\frac{T}{T_0} + Nk\ln\frac{V}{V_0}\Big) \\ &= Nk\,\Delta\ln\frac{V}{V_0} \\ &= Nk\Big(\ln\frac{2V}{V_0} - \ln\frac{V}{V_0}\Big) \\ &= Nk\ln 2 \end{aligned}$$

This is exactly what we get from $S = k\ln\mathcal{N}$: doubling the volume of the gas doubles the number of possible locations of each of its N molecules, so that the number of microscopic states increases by a factor of 2^N and [50]

$$\Delta S = k\ln\Big(2^N\mathcal{N}\Big) - k\ln\mathcal{N} = k\ln 2^N = Nk\ln 2 \quad \checkmark$$

If instead we double the temperature T of the gas while keeping its volume constant, the change in the thermodynamic entropy is

$$\Delta S = \Delta\Big(\tfrac{1}{2}DNk\ln\frac{T}{T_0} + Nk\ln\frac{V}{V_0}\Big)$$

[50] In fact, if we go back to eq. (13.74) of §13.8.1, $S = (\text{const})\ln\mathcal{N}$, matching up the statistical mechanical ΔS with the thermodynamic ΔS requires, as we had then claimed, that the constant in front of the logarithm be the Boltzmann constant k.

$$
\begin{aligned}
&= \tfrac{1}{2} DNk \, \Delta \ln \frac{T}{T_0} \\
&= \tfrac{1}{2} DNk \Big(\ln \frac{2T}{T_0} - \ln \frac{T}{T_0} \Big) \\
&= \tfrac{1}{2} DNk \ln 2
\end{aligned}
$$

Matching this up with $S = k \ln \mathcal{N}$ is a bit more involved than it was for volume changes: If we look at the Boltzmann factor corresponding to the contribution of the v_x degree of freedom to a molecule's energy,

$$e^{-E/kT} = e^{-\frac{1}{2}mv_x^2/kT}$$

we see that $v_x^2 \sim T$, so that $v_x \sim \sqrt{T}$. Doubling the temperature will in a sense therefore increase the range of values of v_x by a factor of $\sqrt{2}$. Since there is such a factor of $\sqrt{2}$ increase for each degree of freedom of each molecule, the number of possible states of the system will increase by a factor of

$$\Big(\sqrt{2}\Big)^{D \cdot N} = 2^{\frac{1}{2}DN}$$

This will translate into a change in entropy of

$$\Delta S = k \ln\Big(2^{\frac{1}{2}DN} \mathcal{N}\Big) - k \ln \mathcal{N} = k \ln 2^{\frac{1}{2}DN} = \tfrac{1}{2} DNk \ln 2 \quad \checkmark$$

Note that we have matched only changes in entropy. The value of the entropy, like that of energy, can be shifted by an arbitrary constant: just as what mattered physically was the change in energy along the trajectory of a system — or, more precisely, the lack of change of energy, as dictated by conservation of energy — , what matters physically is the change in entropy as a system is brought by a process from one state to another: the second law dictates that this change in entropy is nonnegative.

Note also that the increase $Nk \ln 2$ in entropy when the volume is doubled is the same whether the doubling occurs during a reversible isothermal expansion or an irreversible removal of a membrane confining the gas to one half of a box. In the latter case, nothing other than the gas has undergone any change, so that the total entropy of the universe has increased by $Nk \ln 2$: irreversible processes increase total entropy. In terms of probability, this increase in entropy corresponds to the gas's having migrated, after removal of the membrane, from a state of being confined to one half of the box to the 2^N times more likely state of being evenly distributed throughout the whole box. What makes this process irreversible is the extreme unlikeliness of the molecules all spontaneously returning to the half of the box that originally contained them — an occurrence with a probability of only 2^{-N}.

If the doubling in volume is instead the result of a reversible isothermal expansion in a reservoir at the same temperature T as the gas, then the

corresponding change ΔS_r in the entropy of the reservoir is

$$\Delta S_r = \int \frac{dQ}{T} = \frac{Q_r}{T} = -\frac{Q_g}{T}$$

where Q_r is the heat absorbed by the reservoir and Q_g the heat absorbed by the gas: since the heat absorbed by the gas is the heat shed by the reservoir, $Q_r = -Q_g$. And by eq. (13.30b), the heat absorbed by the gas as it expands from V to $2V$ is

$$Q_g = pV \ln \frac{2V}{V} = pV \ln 2 = NkT \ln 2$$

The change in the entropy of the reservoir is therefore

$$\Delta S_r = -\frac{Q_g}{T} = -\frac{NkT \ln 2}{T} = -Nk \ln 2$$

The decrease in the entropy of the reservoir thus exactly cancels the increase in the entropy of the gas: there is no change in total entropy during a reversible process. We have simply shifted entropy from the reservoir to the gas, and can shift it back by reversing the process.

Were we to accomplish the expansion by removing a membrane and then try to return the gas to its original state by an isothermal compression, the change in the entropy of the gas would be reversed, a decrease of $-Nk \ln 2$. But the work we would do to compress the gas would result in heat $NkT \ln 2$ being dumped into the reservoir, with the result that its entropy would rise by

$$\frac{NkT \ln 2}{T} = Nk \ln 2$$

The $Nk \ln 2$ increase in entropy during the irreversible expansion is here to stay.

Note, however, that although it might seem otherwise from reading the newspaper, the second law does not mean that *everything* always becomes more disordered and descends toward anarchy and chaos. As we just saw, the entropy of parts of a larger system can decrease; the second law mandates only that the total change be nonnegative. Your hair, for example, tends to become progressively more disheveled over the course of a day, but you can always restore it order with a comb or brush. The second law does not prohibit grooming yourself. It merely dictates that any decrease in the entropy of your personal appearance be offset by an increase in entropy elsewhere that is at least as large. In the case of coiffure, the heat energy produced by the physiological work of combing or brushing results in an increase in the temperature and entropy of your body and the surroundings.

As you are probably aware even from your own as yet rather brief life experience, humans are separated from the lower animals principally by their

egotism and capacity for irrationality: people do not let even the most palpably indisputable facts stand in the way of what they would like to believe, and what they would like to believe most is that there is something very special and exalted about themselves. That the Earth is not the center of the universe, that humans are just another variety of animal and share a common ancestor with apes, that most of what goes on in the human mind is not only unconscious but governed by the most base and primitive desires — this does not sit well with a humanity that childishly views itself as the whole purpose of the universe's existence. In the present context, one of the favorite arguments against Darwin's theory of evolution is that it violates the second law of thermodynamics: their reasoning is that the spontaneous evolution of life from the chaos of a primordial soup to biological molecules to single-celled organisms and ultimately to the "higher" forms of life constitutes a steady increase in order and therefore a forbidden decrease in entropy. What rubbish. By this kind of reasoning, you would never be able to comb your hair, because doing so would constitute a decrease in entropy. What cannot decrease is total entropy, and the chemical reactions by which biological molecules were formed in the primordial soup occurred for the same reason as any other chemical reactions: because the formation of the new molecules brought the system of molecules to a lower energy state. As a result of this going "downhill" in potential energy, heat is created — thereby raising the entropy of the surroundings. In spite of the decrease in entropy from the increased order in the molecular organization, the total entropy of the universe actually increases. So the theory of evolution is entirely consistent with the second law of thermodynamics.[51]

[51] And so also is the perpetuation of life, as long as the system consisting of the Earth and its inhabitants continues to receive from the Sun energy that can do the work required for the reduction of entropy associated with life.

13.9 Problems

Some data for water that you will almost certainly find useful:

Latent heat of fusion	79.6 cal/g	= 333 J/g
Latent heat of vaporization	539 cal/g	= 2260 J/g
Specific heat (liquid)	1 cal/g°C	= 4.186 J/g°C
Specific heat (ice)	0.48 cal/g°C	= 2.01 J/g°C
Specific heat (steam)	0.48 cal/g°C	= 2.01 J/g°C

1. (a) Is there a temperature at which the degrees Fahrenheit will equal the degrees Celsius? If so, what is this temperature?

   (b) Is there a temperature at which the degrees Fahrenheit will equal the degrees Kelvin? If so, what is this temperature?

   (c) Is there a temperature at which the degrees Celsius will equal the degrees Kelvin? If so, what is this temperature?

   (d) Is there any point to these stupid questions? If so, what is this point?

2. You drop a tomato from rest from the roof of a three-storey (10 m high) building. How much does the temperature of the tomato rise upon impact with the pavement? Treat the tomato as essentially water — as most of those monstrosities you find in supermarkets these days in fact are.

   You should find that the temperature rise is minuscule, which is a good thing: if mechanical energies of everyday magnitude resulted in very large temperature rises, people would be going up in flames in food fights, and whenever you ran you'd have to be careful to slow down gradually for fear of self-immolation. "Live fast, die young" would take on a whole new meaning.

3. At what speed would you have to hurl a 0°C ice cube pointblank at the back of a younger sibling's head in order that it melt completely on impact?

4. A coffee pot with a 500 W heating element is used to boil 1.0 liter of water originally at room temperature (20°C).

   (a) How long does it take to bring the water to a boil?

   (b) How long would it subsequently take to boil off the water completely?

   (c) How high above the ground could a standard 5.0 kg test cat be raised with the energy needed to bring the water to a boil?

   (d) Would it be more energy-efficient to boil the water and dump it on the cat or to drop the cat from the roof of a tall building?

5. If you are in decent shape, you can burn about 800 food calories per hour during moderate aerobic exercise, with roughly 75% of these calories lost as waste heat inside your body.

   (a) If you dissipate all of your waste heat by evaporating sweat, how much do you sweat each hour? For simplicity, treat the sweat as though it were pure water rather than salt water. Also note that evaporation of course occurs at body temperature; it would be very unpleasant if your sweat had to be at the boiling point before evaporating.

   (b) In what other ways do you dissipate body heat?

6. The average man burns off 2200 food calories per day, the average woman, 1800. This difference is sometimes cited as evidence that women have slower metabolic rates than men. Average height for men and women is, however, $5'9''$ and $5'4''$ (1.75 and 1.6 m), respectively. Show that difference in size rather than metabolic rate can therefore entirely account for the difference in calories burned.

7. As a ballpark figure, let's take 2000 food calories per day as the background metabolic rate of the typical person.

   (a) 16 students take an hour-long nap during class. How much heat do they give off?

   (b) Estimate the volume of air in the classroom.

   (c) If the ventilation system is malfunctioning as usual and all of the body heat goes into the air, how much does the temperature in the room rise? The density and specific heat of air are $1.29\,\mathrm{kg/m^3}$ and $1020\,\mathrm{J/kg°C}$. You may assume that the teacher is coldblooded.

   (d) In what ways is this calculation unrealistic?

8. If the humidity (a.k.a. dew point) is high enough, a glass of an iced drink — a mixture of ice and liquid water at 0°C — will "sweat." What thermodynamic effect does this sweating have on the iced drink?

9. (Based on the true story of the dead mouse in the dishpan from the glorious bachelor days of a friend of the author.)[52] Stricken with a sudden yearning for buffalo wings, you heat 3.5 kg of 10W-30 motor oil to 200°C in a metal wastebasket, turn off the heat, and dump in the frozen remains of a 2.0 kg, −10°C dead rat that you found in your dishwater. The oil has a specific heat of 0.70 cal/g°C, and for simplicity you may treat the rat as essentially water (though of course solid water at −10°C).

   (a) Solve for the equilibrium temperature.

   (b) Would this yield satisfactory culinary results?

   (c) How is food actually deep-fried? How does deep-frying differ from baking? And why is everything that tastes good always bad for you?

10. A cute, furry little kitten is stuffed into a hermetically sealed flask fitted with a piston. Describe a process by which you could bake the kitten

    (a) Isochorically.

    (b) Isobarically.

    (c) Isothermally.

    (d) Adiabatically.

    In each case, explain how the pressure and volume of the gas surrounding the kitten would change as the kitten was baked, and whether the gas would do work or have work done on it.

11. If you hold your mouth open and blow on your hand, your breath feels warm. If you pucker and blow, as when cooling hot food, your breath is relatively cool. Explain this. See the footnote if you need a hint.[53]

    This same effect is seen in reverse in a bicycle pump: after several quick pumps, the base of the pump will be warm to the touch.

[52] This friend was not known for his housekeeping — you didn't dare take your shoes off and walk around his apartment unless your tetanus shot was up to date. And while dishes were washed, filling the dishpan with fresh soapy water every day was regarded as a waste of soap, hot water, and effort — once a week was deemed often enough. One day a mouse, evidently either trying to steal a drink or overcome by the fumes, fell into the dishpan and, as it turns out, could tread water only for so long. When he next went to do the dishes and was rummaging around in the murky water for the dishrag, said friend was surprised to discover that not everything that feels like a dishrag is in fact a dishrag.

This incident was followed by the story of the mouse traps laid out throughout the apartment. Immediately before leaving on a two-week trip. In July.

Fortunately this friend has since been housebroken.

[53] Think in terms of $dW = p\,dV$ and remember that the breath leaving your mouth is expanding into the outside air very quickly.

12. A volume of gas starts at pressure p_0 and volume V_0. Initially, four chemists are in the room with the gas. As the gas is heated isobarically until its volume doubles, half of the chemists leave the room and are replaced by twice as many biologists as there were chemists initially. As the gas is subsequently cooled isochorically to half the pressure, as many biologists leave the room as there are chemists in the room. The gas is then further cooled, isobarically, until it returns to its original volume, during which interval the number of biologists in the room is decreased by half the number of chemists who have left the room. How much work has the gas done overall, and what are all those biologists and chemists doing there?

13. An ideal gas is heated isobarically from T_1 to T_2. Show that the work done by the gas is
$$W = Nk(T_2 - T_1)$$

14. Recall that the internal energy of an ideal monatomic gas is $U = \frac{3}{2}NkT$, so that $dU = \frac{3}{2}Nk\,dT$. Determine, in terms of dT, the corresponding heat dQ absorbed and work dW done during

    (a) An isochoric process.

    (b) An isobaric process.

    (c) An isothermal process.

    (d) An adiabatic process.

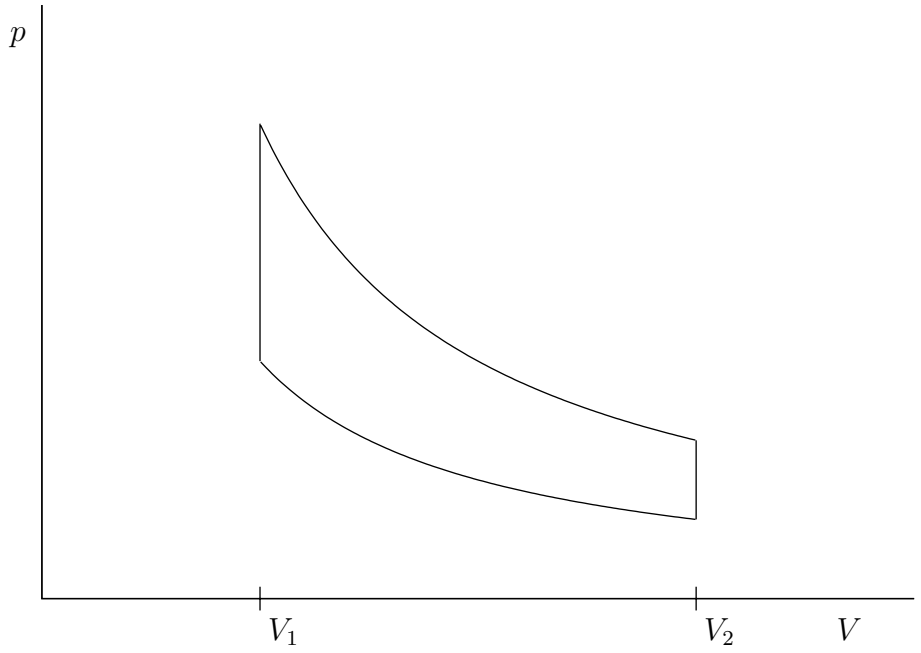


Figure 13.8: Problem 15

15. A sample of an ideal monatomic gas of N molecules is put through the cycle shown in fig. (13.8), which consists of two isochoric and two isothermal legs. The isochoric processes take place at volumes V_1 and V_2 and the isothermal processes at temperatures T_C and $T_H > T_C$.

    (a) Which of the isothermal curves corresponds to the higher temperature T_H?

    (b) Determine the ΔU, Q, and W for each leg when the cycle is run as a heat engine.

    (c) Determine the efficiency of this cycle when it is run as a heat engine.

    (d) Determine the efficiency of this cycle when it is run as a refrigerator.

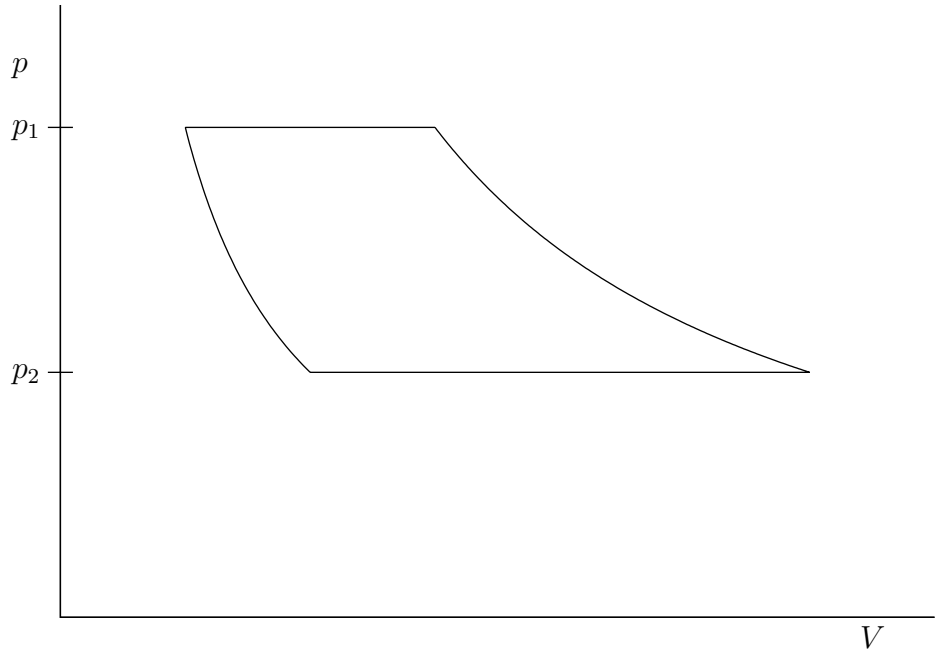


Figure 13.9: Problem 16

16. A sample of an ideal monatomic gas of N molecules is put through the cycle shown in fig. (13.8), which consists of two isobaric and two isothermal legs. The isobaric processes take place at pressures p_1 and p_2 and the isothermal processes at temperatures T_C and $T_H > T_C$.

    (a) Which of the isothermal curves corresponds to the higher temperature T_H?

    (b) Determine the ΔU, Q, and W for each leg when the cycle is run as a heat engine.

    (c) Determine the efficiency of this cycle when it is run as a heat engine.

    (d) Determine the efficiency of this cycle when it is run as a refrigerator.

17. Suppose the condenser of your refrigerator-freezer draws 1500 W of electrical power.

    (a) If the freezer compartment is kept at −10°C and the temperature in the surrounding room is 20°C, what is the theoretical limit on the efficiency of the freezer?

    (b) How much heat energy must be removed from 500 g of 20°C water to turn it into ice? (We're just worried about turning the water to ice here, not lowering it all the way to −10°C.)

    (c) How much heat energy would your freezer dump into the room during this process?

    (d) How quickly could your freezer in principle turn this water into ice?

    (e) What practical considerations and physical effects have we neglected that greatly lengthen the actual time needed to freeze the water?

18. Heating your home by running electricity through an electric base-ray heater or by burning oil or gas is in principle 100% efficient, although some of the heat does go up and out the chimney when you burn oil or gas.[54] It is, however, possible to do better than 100%: some homes are equipped with central air conditioning that can be run backward during the winter as a heat pump. In the summer, the unit takes heat out of the house and dumps it outside; in the winter, that same refrigeration cycle is used in reverse, to take heat from outside and dump it in the house.

    Suppose that during the heating season the temperature is typically 68°F in your house and 40°F outside.[55] If the heat pump ran like a Carnot refrigerator, how much heat would be dumped into the house for every Joule of electrical energy used to run the pump?

[54] Conserving energy by turning off the lights when you leave a room has become a compulsively reflexive action for some people, but if your house is heated electrically, during the heating season you might as well leave all your incandescent lights blazing away like the Sun: the vast majority of the electrical power consumed by incandescent lights goes into heat. Since you have to heat your home, and since that heat is going to be produced by consuming electrical power anyway, you might as well heat it in a way that doesn't leave you blundering around in the dark. The same would apply to other indoor electrical appliances during the heating season — computers, televisions, electric chainsaws, etc.

Of course, during the cooling season it is advantageous to minimize your use of electric appliances: the heat they produce during the cooling season of no benefit, and you then have to use even more energy to remove this waste heat by air conditioning.

[55] In principle heat pumps would work anywhere that the outside temperature is above absolute zero, but unfortunately practical engineering limitations restrict their use to areas where the winters are relatively mild. And even in areas where heat pumps are effective, they are of course not nearly as efficient as Carnot refrigerators.

19. (a) We hear tell that in the early days of the transition from the ice box, which was cooled by blocks of ice brought by the ice man, to the larger modern electric refrigerator, sometimes, for want of any better space, the refrigerator would end up in a closet or in a small, closed pantry.

    i. People were mystified to find that the refrigerator did not cool effectively when run inside a closet. Explain this. (Meaning, explain why the refrigerator wouldn't run well in a closet, not why people were mystified. *Duh!*)

    ii. Boring sizeable holes, one set at the top of the closet door and another set at the bottom, would enable the refrigerator to run reasonably well. Explain why this trick worked.

    (b) We have noted that because of the need for a place to dump the heat, it is impossible to design an in-room air conditioner, that is, a standalone unit contained entirely within the room being cooled. You may, however, have seen ads for standalone units that are a combination air conditioner and dehumidifier. How might this combination work?

20. We have seen that the efficiency of a Carnot engine is

$$\eta = 1 - \frac{T_C}{T_H}$$

Other kinds of heat engines, although the expressions for their efficiencies are different, also tend to be more efficient when operated between greater temperature differences.

Does this mean that people who live in colder climates have more efficient metabolisms?

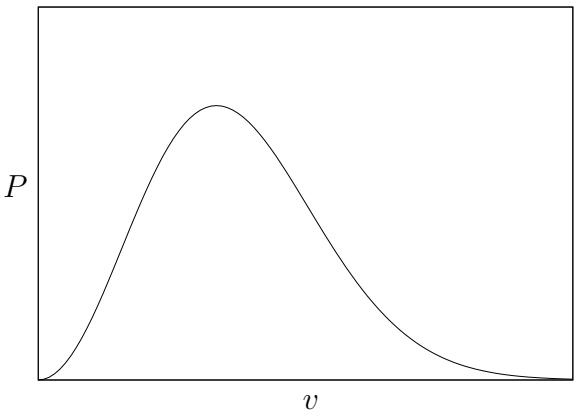


Figure 13.10: Problem 21

21. Recall that in statistical mechanics the probability of being in a state of energy E is proportional to the Boltzmann factor $e^{-E/kT}$.

(a) If the molecules of an ideal gas each have mass m and their energies are entirely of the translational kinetic form

$$\tfrac{1}{2}mv^2 = \tfrac{1}{2}m(v_x^2 + v_y^2 + v_z^2)$$

show that the probability dP of a molecule's speed being in the range v to $v + dv$ is given by

$$\frac{dP}{dv} = 4\pi\left(\frac{m}{2\pi kT}\right)^{\frac{3}{2}} v^2\, e^{-mv^2/2kT} \tag{13.78}$$

that is, that the probability of a molecule's speed being between v_1 and v_2 is

$$P_{v_1 \text{ to } v_2} = 4\pi\left(\frac{m}{2\pi kT}\right)^{\frac{3}{2}} \int_{v_1}^{v_2} dv\, v^2\, e^{-mv^2/2kT}$$

Note that since v represents a molecule's speed, you will want to work in spherical coordinates, for which, the distribution of three-dimensional velocities being spherically symmetric,[56]

$$dv_x\, dv_y\, dv_z = 4\pi v^2\, dv$$

Also remember that your probabilities must be normalized, so that the probability of a molecule's speed being between zero and infinity is unity. Eq. (13.78) is known as the Maxwell-Boltzmann distribution and is plotted in fig. (13.10).

(b) How does this distribution of velocities help to explain evaporation?

[56] This is just

$$dx\, dy\, dz = 4\pi r^2\, dr$$

expressed for a velocity vector rather than a coordinate vector.

22. Recall that in statistical mechanics the probability of being in a state of energy E is proportional to the Boltzmann factor $e^{-E/kT}$.

    Suppose we have an ideal monatomic gas at a uniform temperature T within a vertical column. Each molecule has mass m, and the bottom of the column is at height $h = 0$.

    Using the technique of pp.680ff.,

    (a) Show that the average gravitational potential energy of a molecule is

    $$\overline{U}_{grav} = kT$$

    (b) Show that the average height of a molecule is

    $$\overline{h} = \frac{kT}{mg}$$

    (c) How will the density of the gas vary as a function of height?

13.10 Sketchy Answers

(1a) -40.

(1b) 574.

(1d) No, not really.

(2) 0.023°C.

(3) 816 m/s.

(4a) 11 min.

(4b) 75 min.

(4c) 6.8×10^3 m.

(5a) 1.1 kg.

(7a) 1.3×10^6 cal.

(7c) 4.24×10^3 °C·m^3 divided by your result for the previous part.

(9a) 72°C.

(12) $\frac{1}{2}p_0V_0$. And for those of you who bothered to figure out what was never asked, 2 chemists and 5 biologists.

(14c) *Hello!*

(15b) Your nonzero answers should include

$$\pm\tfrac{3}{2}Nk(T_H - T_C), \quad NkT_H \ln\frac{V_2}{V_1}, \quad -NkT_C \ln\frac{V_2}{V_1}$$

(15c) $$\frac{(T_H - T_C)\ln\frac{V_2}{V_1}}{T_H \ln\frac{V_2}{V_1} + \frac{3}{2}(T_H - T_C)}$$

(15d) $$\frac{T_C \ln\frac{V_2}{V_1} + \frac{3}{2}(T_H - T_C)}{(T_H - T_C)\ln\frac{V_2}{V_1}}$$

(16b) Your nonzero answers should include

$$\pm Nk(T_H - T_C), \quad \pm\tfrac{3}{2}Nk(T_H - T_C), \quad \pm\tfrac{5}{2}Nk(T_H - T_C),$$
$$NkT_H \ln\frac{p_2}{p_1}, \quad -NkT_C \ln\frac{p_2}{p_1}$$

(16c) $$\frac{(T_H - T_C)\ln\frac{p_2}{p_1}}{T_H \ln\frac{p_2}{p_1} + \frac{5}{2}(T_H - T_C)}$$

(16d) $\dfrac{T_C \ln \frac{p_2}{p_1} + \frac{5}{2}(T_H - T_C)}{(T_H - T_C) \ln \frac{p_2}{p_1}}$

(17a) 8.77.

(17b) 2.1×10^5 J.

(17c) 2.3×10^5 J.

(17d) 16 sec.

(18) 11 J.

Part IV

Electromagnetism for Big People

Chapter 14

The Maxwell Equations: An Overview

14.1 The Maxwell Equations

Electromagnetism is a beautiful subject:[1] all of the great diversity of electromagnetic phenomena are governed by just four relations known as the Maxwell equations.[2] The purpose of this chapter is to briefly introduce you to the Maxwell equations; developing a fuller, working understanding of their import and applications will be the work of later chapters. The hope is that, though you may not immediately grasp everything in this chapter, you will leave it with an abiding awareness of the unity of electromagnetism and will, as we delve into details and specific applications in the following chapters, not lose sight of the fact that all of the great variety of electromagnetic phenomena we discuss are just special cases and consequences of the Maxwell equations.[3]

[1] Which is good, because it certainly isn't very fertile ground for humor. There just isn't a whole lot to work with. "So there were these two electrons in a bar, see, and one ...."

[2] At a higher level of formalism, electromagnetism is in fact governed by a single four-vector relation, $\partial_\mu F^{\mu\nu} = j^\nu$, where j^ν is the four-vector current, $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$ is the field strength tensor, and A^μ is the four-potential. And at a still higher level (as we will show in detail in §22.6.1), electromagnetism can be derived from the symmetry of the unit circle. You don't get much simpler than that.

The neatness and simplicity of electromagnetism are why Einstein formulated relativity theory, not in terms of mechanics — an ugly, messy, sprawling subject that ill reflects the true fundamental principles of nature — , but in terms of electromagnetism: his famous 1905 first paper on relativity in *Annalen der Physik* was "Zur Elektrodynamik bewegter Körper" (On the Electrodynamics of Moving Bodies).

[3] This is very contrary to the approach taken by most introductory textbooks, which is to spend a lot of time talking about people sending keys up on kites, chasing one magnet with another around coffee tables, etc., introducing you to electromagnetic phenomena in such a messy, disjointed, piecemeal way that there is no hope of your ever seeing the forest

The Maxwell equations can be expressed in two forms, a differential form involving divergences and curls and, by means of Gauss's and Stokes's theorems, an equivalent integral form involving surface and line integrals. In their differential form, the Maxwell equations are

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\,\rho \tag{14.1a}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{14.1b}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{14.1c}$$

$$\nabla \times \mathbf{B} = \mu_0\,\mathbf{j} + \epsilon_0\mu_0\frac{\partial \mathbf{E}}{\partial t} \tag{14.1d}$$

and in their integral form they are

$$\oint_S dA\;\hat{\mathbf{n}} \cdot \mathbf{E} = \frac{1}{\epsilon_0}\,q_{enclosed} \tag{14.2a}$$

$$\oint_S dA\;\hat{\mathbf{n}} \cdot \mathbf{B} = 0 \tag{14.2b}$$

$$\oint_C \mathbf{E} \cdot d\mathbf{r} = -\frac{d\Phi_B}{dt} \tag{14.2c}$$

$$\oint_C \mathbf{B} \cdot d\mathbf{r} = \mu_0\,I_{enclosed} + \epsilon_0\mu_0\frac{d\Phi_E}{dt} \tag{14.2d}$$

Probably this all looks like something out of the Voynich manuscipt. But please do not freak out. If you find yourself hyperventilating, try breathing into a paper bag until you calm down. In the following sections, as we discuss the meanings of these equations and the physical quantities and concepts involved in them, you will see that there is nothing to be afraid of, and as you work with the Maxwell equations in later chapters you will not only grow comfortable with them, but even come to think of them as friends.[4]

14.2 Charge & Current

Unless you have led a very sheltered life, you know that electric charge can be either positive or negative. The conventional symbol for charge is q. Maybe

through the trees. Historical vignettes and engineering trivia are entertaining to read later on, after you already have an understanding of electromagnetism, but they only obscure the issues when you are first learning the subject. While you may find the inevitable abstractness of our approach challenging in some respects, at least you will not lose sight of the big picture.

[4]Then again, people tell you all kinds of ****, so who knows?

not the first symbol you would have thought of, but c is already taken by the speed of light, and at any rate if you were to use c for charge, what would you use for current? Gotcha there.

Anyway, there are more than 4 bazillion systems of units for electromagnetic quantities. In order to be consistent with the vast majority of introductory physics texts and with the usage of engineers, we will be using one of the least sensible systems, the rationalized MKSA system.[5] In this system, the unit of electric charge is the Coulomb (C). If you are talking about isolated electric charge, a Coulomb, as it turns out, is a very large quantity of charge — about what you'd find in a typical backyard-variety lightning bolt.

When we are dealing with a continuous distribution of electric charge, it is convenient to work with charge per unit volume, which is known as the charge density and, by analogy to mass densities, conventionally denoted by the symbol ρ:[6]

$$\rho = \frac{dq}{dV} \tag{14.3}$$

Electric current is just the motion of electric charge. Specifically, the current I through a surface S is defined as the rate at which charge is passing through that surface:

$$I = \frac{dq}{dt} \tag{14.4}$$

where dq is the amount of charge that passes through the surface S during the time interval dt. Our unit for current is the Ampère or amp (A): 1 A = 1 C/sec.

From this definition, we can also derive an expression for the rate at which charge is flowing out of or into individual points by looking at the flow through infinitesimal patches of a surface. This point-like flow of charge, called the current density, is conventionally denoted by $\mathbf{j}$ — boldface, because it is a vector that points in the direction of motion of the charge.[7] If dq is the amount of charge that passes through the infinitesimal patch dA in time dt, we can use eq. (14.3) to re-express dq as

$$dq = \frac{dq}{dV}\, dV = \rho\, dV$$

[5] Peer pressure is a terrible thing.

In case you were curious, the *MKSA* stands for *meter-kilogram-second-amp*. The *rationalized* part apparently refers to the wishful thinking that there is virtue in this system of units, when in fact it is just the qwerty keyboard of electromagnetism.

[6] We will also eventually be dealing with linear charge densities λ and surface charge densities σ, but we'll tackle those when the need arises; volume densities are sufficient for our present discussion.

[7] Actually, for negative charge $\mathbf{j}$ will point in the opposite direction — as we will see later, negative charge moving one way is equivalent to an equal quantity of positive charge moving diametrically the opposite way.

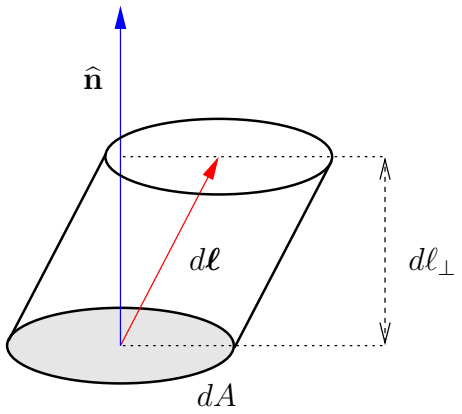


Figure 14.1: Life Can Be Complicated

where dV is the volume of charge that moves through the patch dA. If $d\boldsymbol{\ell}$ is the displacement of the charge dq during dt, it might be tempting to write $dV = dA\, d\ell$, but we must take into account that in general dq will not be moving perpendicularly through the patch dA. As you can see from fig. (14.1), if $\hat{\mathbf{n}}$ is the normal to the patch dA and $d\boldsymbol{\ell}$ the displacement of the charge dq passing through dA, the cylindrical volume dV is actually

$$dV = dA\, d\ell_\perp = dA\, \hat{\mathbf{n}} \cdot d\boldsymbol{\ell}$$

Our expression for dq thus becomes

$$dq = \rho\, dV = \rho\, dA\, \hat{\mathbf{n}} \cdot d\boldsymbol{\ell}$$

so that the infinitesimal current dI that flows through the patch dA is, by eq. (14.4),

$$dI = \frac{dq}{dt} = \frac{\rho\, dA\, \hat{\mathbf{n}} \cdot d\boldsymbol{\ell}}{dt} = \rho\, dA\, \hat{\mathbf{n}} \cdot \frac{d\boldsymbol{\ell}}{dt} = \rho\, dA\, \hat{\mathbf{n}} \cdot \mathbf{v}$$

where we have noted that the displacement $d\boldsymbol{\ell}$ of the charge dq divided by dt is just the velocity $\mathbf{v}$ of dq. Integrating this to get the total current I through the whole surface S and putting the ρ next to the $\mathbf{v}$, we arrive at

$$I = \int_S dA\, \hat{\mathbf{n}} \cdot (\rho \mathbf{v})$$

This leads us to define the current density $\mathbf{j}$ at each point on the surface as

$$\mathbf{j} = \rho \mathbf{v} \tag{14.5}$$

The current I flowing through a surface S is then just the integral of the normal component of the current density over the surface:

$$I = \int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{j} \tag{14.6}$$

Again, note that only the component of **j** perpendicular to the surface will contribute to the flow of charge through it; to the extent that **j** is parallel to the surface, the charge will be flowing by the surface rather than through it.

14.3 Gauss's Law & Electric Fields & Forces

The first of the Maxwell equations (14.1) is known as Gauss's law:[8]

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\rho \tag{14.1a}$$

where ρ is the charge density introduced in the previous section,

$$\epsilon_0 = 8.854187817 \times 10^{-12} \frac{\mathrm{C}^2}{\mathrm{N\,m}^2}$$

is a physical constant known as the electric permittivity of the vacuum,[9] and the other new animal, **E**, is known as the electric field. Note that **E** is a vector field: it has both a magnitude and a direction at every point in space and time.

As we saw in §2.5, the divergence of a vector field gives a measure of how much field is diverging from (or, if negative, converging into) the point in space at which the divergence is being calculated. Gauss's law is therefore telling us that electric charge is the source (or, in the case of negative charge, the sink) of electric field: to the extent that there is a positive charge density ρ at a point, there is electric field emanating from that point; to the extent that there is a negative charge density at a point, there is electric field converging into and being, as it were, swallowed up by it. And this electric field is not just an abstraction; as we will see in §16.5.1, it has an energy associated with it and is thus a living, breathing physical animal in its own right.

That doesn't make it any less obscure an animal, of course. At this very moment you may be thinking, But what, pray tell, doth signify this electric field whereof thou speakest? Or perhaps even, Dude, what the **** are you talking about? And the answer to either question would be this: the electric field is the direct cause of electric forces; it is the vehicle by which electric forces are conveyed. To wit, the electric force $\mathbf{F}_{electric}$ on a point charge q is given by

$$\mathbf{F}_{electric} = q\mathbf{E} \tag{14.7}$$

where **E** is the electric field *at the location of* q.

[8] Henceforward, you have to be careful to distinguish between Gauss's law (the physical law (14.1a)) and Gauss's theorem (the mathematical theorem (2.6)).

[9] The value of the electric permittivity is different in different substances; the 0 subscript conventionally indicates that the "substance" is the vacuum. In more sensible systems of units, $\epsilon_0 = 1/4\pi$.

The communication of an electric force between two charges — say, X and Y — is thus a twofold process: Charge X gives rise to an electric field that permeates all of spacetime. At any given time, it is the value of that electric field at the location of charge Y that gives rise to the electric force felt by Y. There is therefore no problem of action at a distance, like the instantaneous communication of the Newtonian gravitational force between masses separated by a finite distance: the force felt by Y depends, not on where X is at that instant, but on the electric field *at the location of* Y at that instant. Moving X will alter the force felt by Y, but in a way entirely consistent with relativity: as we will see in §14.9, the Maxwell equations dictate that changes in the electric field of X propagate outward from X, not instantaneously, but at a finite speed — the speed of light.

From eq. (14.7), you can see that our units for the electric field will be those of force/charge: N/C. According to Gauss's law (14.1a), the electric field to which the charge density ρ gives rise will involve a factor of ϵ_0, which will therefore also enter into our force relation (14.7). For example, as we will see in §15.2, the relation for the force between two point charges q_1 and q_2 that are separated by a distance r is

$$F = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r^2}$$

The electric permittivity ϵ_0 may thus be regarded as a sort of conversion factor that yields a force in Newtons between charges measured in Coulombs.

Eq. (14.1a) is the differential form of Gauss's law: it applies at each point in space and time. The equivalent integral form is

$$\oint_S dA\ \hat{\mathbf{n}} \cdot \mathbf{E} = \frac{1}{\epsilon_0}\, q_{enclosed} \tag{14.2a}$$

where $q_{enclosed}$ is the net charge enclosed by the closed surface S. To see that the differential and integral forms of Gauss's law are indeed equivalent, we first show that the integral form follows from the differential form if we apply Gauss's theorem (2.6) to the electric field:

$$\oint_S dA\ \hat{\mathbf{n}} \cdot \mathbf{E} = \int_V dV\ \nabla \cdot \mathbf{E} \tag{14.8}$$

By integrating the normal component of both sides of the differential form of Gauss's law over a volume V, we can express the right-hand side of this relation in terms of the charge density:

$$\int_V dV\ \nabla \cdot \mathbf{E} = \int_V dV\ \frac{1}{\epsilon_0}\,\rho \tag{14.9}$$

Using the definition (14.3) of charge density, we can further gerrymander the right-hand side into

$$\int_V dV\ \frac{1}{\epsilon_0}\,\rho = \int_V dV\ \frac{1}{\epsilon_0}\,\frac{dq}{dV} = \frac{1}{\epsilon_0}\int_V dq = \frac{1}{\epsilon_0}\, q_{enclosed} \tag{14.10}$$

where we have noted that the sum over all the infinitesimal charges dq within the volume V enclosed by the surface S is just the total charge $q_{enclosed}$ enclosed by that surface. Putting together eqq. (14.8)-(14.10), we arrive at the integral form (14.2a) of Gauss's law.

This proof is also reversible: if we start with the integral form of Gauss's law, we can again use Gauss's theorem and the definition of charge density to derive its differential form:

$$\begin{aligned}
\oint_S dA\, \hat{\mathbf{n}} \cdot \mathbf{E} &= \frac{1}{\epsilon_0}\, q_{enclosed} \\
\int_V dV\, \nabla \cdot \mathbf{E} &= \frac{1}{\epsilon_0} \int_V dq \\
&= \int_V dV\, \frac{1}{\epsilon_0}\, \frac{dq}{dV} \\
&= \int_V dV\, \frac{1}{\epsilon_0}\, \rho
\end{aligned}$$

Since this equality must hold for all possible volumes V, the integrands themselves must be equal; otherwise we could always find some V for which the equality would not hold. We therefore conclude that

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\, \rho$$

which is indeed the differential form of Gauss's law.

In the integral form of Gauss's law, S is called the Gaussian surface and, as in Gauss's theorem, the normal $\hat{\mathbf{n}}$ is always the *outward* normal to that surface.

14.4 Magnetic Gauss's Law & Magnetic Fields

The second of the Maxwell equations (14.1) is

$$\nabla \cdot \mathbf{B} = 0 \tag{14.1b}$$

where $\mathbf{B}$ is the magnetic field. This relation doesn't really have a name, but we will informally refer to it as the magnetic Gauss's law. Except that the right-hand side of this magnetic version of Gauss's law is zero rather than something like ρ/ϵ_0, it is exactly like the electric version of Gauss's law. We therefore know, without having to go through the proof again, that this differential form of the magnetic Gauss's law is equivalent to the integral form

$$\oint_S dA\, \hat{\mathbf{n}} \cdot \mathbf{B} = 0 \tag{14.2b}$$

Recall that physically the electric Gauss's law states that electric charge is the source of electric field. Because the right-hand sides of eqq. (14.1b) and (14.2b) vanish, the magnetic Gauss's law is stating that there is no such thing as magnetic charge, that is, that there are no magnetic monopoles.[10] The existence of magnetic monopoles is not only possible theoretically, but desirable: it turns out that their existence would explain why electric charge is quantized. Unfortunately, no experiment to detect them has ever seen one. But if and when a magnetic monopole is seen, the Maxwell equations (14.1b) and (14.1c) for the magnetic field can easily be modified to include the contributions of magnetic charge and current.

One physical consequence of the magnetic Gauss's law is that magnetic fields are at best dipole fields, that is, that magnets and other sources of magnetic field always have both a "north" and a "south" pole: the magnetic field that emanates from the north pole must ultimately curve around back into the south pole; otherwise, there would be a nonzero net contribution to

$$\oint_S dA\, \hat{\mathbf{n}} \cdot \mathbf{B}$$

for surfaces enclosing the magnet.

14.5 Faraday's Law

The third of the Maxwell equations (14.1) is known as Faraday's law:

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{14.1c}$$

Faraday's law tells us that a magnetic field $\mathbf{B}$ that changes as time passes induces, in the sense that it creates or gives rise to, an electric field $\mathbf{E}$ that is a pure curl. This effect is called magnetic induction and is in contrast to the electric field due to electric charge, which is, by Gauss's law, a pure divergence. We'll look into the details and consequences of this in Chapter 18, but for now it is enough simply to note that a changing magnetic field induces an electric field.

Faraday's law may also be expressed in the equivalent integral form

$$\oint_C \mathbf{E} \cdot d\mathbf{r} = -\frac{d\Phi_B}{dt} \tag{14.2c}$$

[10] The term *monopole* is equivalent to *charge* but sounds a lot cooler, which is presumably why people always speak of *magnetic monopoles* rather than *magnetic charge*. In case you were curious, the terminology comes from what is known as the multipole expansion of the electrostatic field in an infinite series of contributions from monopole, dipole, quadrupole, octopole, ..., moments, the monopole moment corresponding to the contribution from net charge.

where C is any closed contour (that is, any closed curve or loop) and Φ_B is the magnetic flux through that loop, with the magnetic flux defined as

$$\Phi_B = \int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{B} \tag{14.11}$$

where S is any surface spanning the loop C. To see that the differential and integral forms of Faraday's law are indeed equivalent, we first show that the integral form follows from the differential form if we apply Stokes's theorem (2.8) to the electric field:

$$\oint_C \mathbf{E} \cdot d\mathbf{r} = \int_S dA\, \hat{\mathbf{n}} \cdot \nabla \times \mathbf{E} \tag{14.12}$$

By integrating the normal component of both sides of the differential form of Faraday's law over a surface S and using the definition (14.11) of magnetic flux, we can gerrymander it into[11]

$$\begin{aligned} \int_S dA\, \hat{\mathbf{n}} \cdot \nabla \times \mathbf{E} &= \int_S dA\, \hat{\mathbf{n}} \cdot \left(-\frac{\partial \mathbf{B}}{\partial t} \right) \\ &= -\frac{d}{dt} \left(\int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{B} \right) \\ &= -\frac{d\Phi_B}{dt} \end{aligned} \tag{14.13}$$

Putting together eqq. (14.12) and (14.13), we arrive at the integral form (14.2c) of Faraday's law.

And if we start with the integral form of Faraday's law, we can again use Stokes's theorem and the definition (14.11) of magnetic flux to reverse the proof and derive its differential form:[12]

$$\oint_C \mathbf{E} \cdot d\mathbf{r} = -\frac{d\Phi_B}{dt}$$

[11] The wary reader will have noticed that a partial derivative with respect to time here very sneakily turns into a total derivative when it is pulled outside of the integral. This change is not gratuitous: the total derivative takes into account that there may be changes in flux not only due to a time-variation in the magnetic field $\mathbf{B}$, but also due to motion of the surface S. And we must allow for such motion in order for the integral form of Faraday's law to be anything like a general relation: even if S is stationary from our perspective, S will be in motion from the perspective of any observer moving relative to us. This will, however, be one of the few places where we will forego the proof of an essential point; the rigorous mathematical justification for the shift from a partial to a total derivative would be, from where we are now, more arduously involved than it would be worth, nor would it offer any illumination beyond the point we have just now made: that the time-variation may be either in the magnetic field or, by virtue of its motion or a change in its shape or extent, in the surface S. Given that relativity is built into the Maxwell equations, it should seem plausible that we should have such a total derivative of the flux to ensure that Faraday's law applies equally well from any observer's perspective. We will say more about this issue in §18.4.

[12] We are of course here pulling the same slippery fast one with total and partial time derivatives as we did above.

$$\int_S dA\,\hat{\mathbf{n}}\cdot\nabla\times\mathbf{E} = -\frac{d}{dt}\left(\int_S dA\,\hat{\mathbf{n}}\cdot\mathbf{B}\right)$$
$$= \int_S dA\,\hat{\mathbf{n}}\cdot\left(-\frac{\partial\mathbf{B}}{\partial t}\right)$$

Since this equality must hold for all possible surfaces S and their normals $\hat{\mathbf{n}}$, the parts of the integrands into which the normal is dotted must themselves be equal; otherwise we could find some surface S for which the equality would not hold. We therefore conclude that

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}$$

which is indeed the differential form of Faraday's law.

As in Stokes's theorem, in the integral form of Faraday's law the normal $\hat{\mathbf{n}}$ to the surface S (and hence the direction we consider to be positive when evaluating the magnetic flux Φ_B) must be taken to be in a right-handed sense relative to the direction in which we are integrating around the loop C.

We would otherwise expect the magnetic flux Φ_B to depend on the particular surface S chosen to span the loop C, but the magnetic Gauss's law (14.2b) dictates that we will get the same value for the flux for any loop that spans C: any closed surface S may be divided into two nonclosed surfaces S_1 and S_2 by drawing a closed loop C on S. The vanishing of the flux coming out of the whole surface S means that the sum of the fluxes coming out of S_1 and S_2 vanishes — or, in other words, that the flux that passes into S_1 is the same as the flux that passes out through S_2. Since this argument applies to any loop and any two spanning surfaces, we may speak simply of the magnetic flux through a loop, without having to specify the surface by which that loop is spanned.

14.6 Ampère's Law & Magnetic Forces

The fourth of the Maxwell equations (14.1) is known as Ampère's law:[13]

$$\nabla\times\mathbf{B} = \mu_0\,\mathbf{j} + \epsilon_0\mu_0\frac{\partial\mathbf{E}}{\partial t} \tag{14.1d}$$

where $\mu_0 = 4\pi\times10^{-7}\,\mathrm{N/A^2}$ is a physical constant known as the magnetic permeability of the vacuum, the magnetic analogue of the electric constant

[13] Strictly speaking, Ampère's law refers to the what eqq. (14.1d) and (14.2d) reduce to when there is no dependence on time:

$$\nabla\times\mathbf{B} = \mu_0\,\mathbf{j} \qquad\qquad \oint_C \mathbf{B}\cdot d\mathbf{r} = \mu_0\,I_{enclosed}$$

But we will take a bit of license and refer to eqq. (14.1d) and (14.2d) as Ampère's law.

ϵ_0.[14] Ampère's law tells us there are two sources of magnetic field: the motion of charge (that is, electric current) and time-variation of the electric field. The first term on the right-hand side, $\mu_0 \mathbf{j}$, is the contribution from the electric current. The second term, $\epsilon_0\mu_0 \partial\mathbf{E}/\partial t$, is an induction contribution analogous to that in Faraday's law, but, since it involves the time-variation of the electric rather than the magnetic field, is referred to as electric induction. This second term is telling us that a time-varying electric field $\mathbf{E}$ induces, again in the sense that it gives rise to or creates, a magnetic field. We'll look into the details and consequences of electric induction in later chapters; for now we will concern ourselves only with the contribution of electric current.

As we will see in Chapter 18, the magnetic field, just like the electric field, turns out to have an energy associated with it and is thus also a living, breathing physical animal. The magnetic field is also the direct cause of magnetic forces: the magnetic force $\mathbf{F}_{magnetic}$ on a point charge q is given by

$$\mathbf{F}_{magnetic} = q\mathbf{v} \times \mathbf{B} \tag{14.14}$$

where $\mathbf{B}$ is the electric field *at the location of* q and $\mathbf{v}$ is the velocity of q. Note that only moving charge gives rise to magnetic field and that in turn only moving charge experiences magnetic force.

As was the case with the electric force, there is no action at a distance: the magnetic force felt by a charge depends on the magnetic field *at that charge's location*, not on the location of the current that gives rise to the magnetic field. Changes in magnetic field, like those in electric field, propagate from their sources at the finite speed of light.[15]

Our units for magnetic field are Tesla (T). Even powerful electromagnets can't produce fields of stronger than a few dozen Tesla; the Earth's rather weak magnetic field is typically on the order of 10^{-5} T.

Ampère's law may also be expressed in the equivalent integral form

$$\oint_C \mathbf{B} \cdot d\mathbf{r} = \mu_0 \, I_{enclosed} + \epsilon_0\mu_0 \frac{d\Phi_E}{dt} \tag{14.2d}$$

where C is any closed contour (loop) and $I_{enclosed}$ and Φ_E are the current and electric flux through that loop, with the electric flux defined as

$$\Phi_E = \int_S dA \, \hat{\mathbf{n}} \cdot \mathbf{E} \tag{14.15}$$

where S is any surface spanning the loop C. To see that the differential and integral forms of Ampère's law are equivalent, we first show that the integral

[14] And as was the case with ϵ_0, the value of the magnetic permeability depends on the substance, with the 0 subscript again conventionally indicating that the "substance" is the vacuum.

[15] In fact, we will see in §14.9 and, more fully, in §22.6.1, that electric and magnetic fields are really the same field; they are just two different manifestations of a single entity, the electromagnetic field.

form follows from the differential form if apply Stokes's theorem (2.8) to the magnetic field:

$$\oint_C \mathbf{B} \cdot d\mathbf{r} = \int_S dA\, \hat{\mathbf{n}} \cdot \nabla \times \mathbf{B} \tag{14.16}$$

If we integrate the normal component of both sides of the differential form (14.1d) of Ampère's law over a surface S and use the definition (14.15) of electric flux and relation (14.6) between the current I and the current density $\mathbf{j}$, we obtain [16]

$$\begin{aligned}\int_S dA\, \hat{\mathbf{n}} \cdot \nabla \times \mathbf{B} &= \int_S dA\, \hat{\mathbf{n}} \cdot (\mu_0 \mathbf{j}) + \int_S dA\, \hat{\mathbf{n}} \cdot \left(\epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t}\right) \\ &= \mu_0 \int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{j} + \epsilon_0 \mu_0 \frac{d}{dt}\left(\int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{E}\right) \\ &= \mu_0 I_{enclosed} + \epsilon_0 \mu_0 \frac{d\Phi_E}{dt}\end{aligned} \tag{14.17}$$

Putting together eqq. (14.16) and (14.17), we arrive at the integral form of Ampère's law. And if we start with the integral form of Ampère's law, we can again use Stokes's theorem, the definition of electric flux, and the relation between current and current density to reverse the proof and derive its differential form:

$$\begin{aligned}\oint_C \mathbf{B} \cdot d\mathbf{r} &= \mu_0 I_{enclosed} + \epsilon_0 \mu_0 \frac{d\Phi_E}{dt} \\ \int_S dA\, \hat{\mathbf{n}} \cdot \nabla \times \mathbf{B} &= \mu_0 \int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{j} + \epsilon_0 \mu_0 \frac{d}{dt}\left(\int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{E}\right) \\ &= \int_S dA\, \hat{\mathbf{n}} \cdot \left(\mu_0 \mathbf{j} + \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t}\right)\end{aligned}$$

Since this equality must hold for all possible surfaces S and their normals $\hat{\mathbf{n}}$, the parts of the integrands into which the normal is dotted must themselves be equal; otherwise we could find some surface S for which the equality would not hold. We therefore conclude that

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} + \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t}$$

which is indeed the differential form of Ampère's law.

As in Stokes's theorem, in the integral form of Ampère's law the normal $\hat{\mathbf{n}}$ to the surface S (and hence the direction we consider to be positive when evaluating both the current and the electric flux through the loop) must be taken to be in a right-handed sense relative to the direction in which we are integrating around the loop C.

[16] Here and below we are pulling the same fast one with partial and total derivatives that we did in the corresponding derivations for Faraday's law, and for the same reasons. (See footnote 11 on p.757.)

14.7 Superposition

One very important property of the Maxwell equations is linearity: they are linear in the electric and magnetic fields and in their sources, the charges and currents. This means that the net electric and magnetic fields to which compound distributions of charge and current give rise are simply the sum (the vector sum, of course) of the electric and magnetic fields due to the various charges and currents making up that compound distribution.

To prove this, suppose that we have two distributions of charge and current: a distribution ρ_1 and $\mathbf{j}_1$ that, by itself, would give rise to electric field $\mathbf{E}_1$ and magnetic field $\mathbf{B}_1$, and a distribution ρ_2 and $\mathbf{j}_2$ that by itself would give rise to electric field $\mathbf{E}_2$ and magnetic field $\mathbf{B}_2$. Distributions 1 and 2 therefore separately obey the Maxwell equations (14.1):

$$\nabla \cdot \mathbf{E}_1 = \frac{1}{\epsilon_0}\,\rho_1 \qquad \nabla \cdot \mathbf{B}_1 = 0$$
$$\nabla \times \mathbf{E}_1 = -\frac{\partial \mathbf{B}_1}{\partial t} \qquad \nabla \times \mathbf{B}_1 = \mu_0\,\mathbf{j}_1 + \epsilon_0\mu_0\frac{\partial \mathbf{E}_1}{\partial t}$$

and

$$\nabla \cdot \mathbf{E}_2 = \frac{1}{\epsilon_0}\,\rho_2 \qquad \nabla \cdot \mathbf{B}_2 = 0$$
$$\nabla \times \mathbf{E}_2 = -\frac{\partial \mathbf{B}_2}{\partial t} \qquad \nabla \times \mathbf{B}_2 = \mu_0\,\mathbf{j}_2 + \epsilon_0\mu_0\frac{\partial \mathbf{E}_2}{\partial t}$$

Adding each of the relations for the second distribution to the corresponding relation for the first, we obtain

$$\nabla \cdot \mathbf{E}_1 + \nabla \cdot \mathbf{E}_2 = \frac{1}{\epsilon_0}\,\rho_1 + \frac{1}{\epsilon_0}\,\rho_2$$
$$\nabla \cdot \mathbf{B}_1 + \nabla \cdot \mathbf{B}_2 = 0$$
$$\nabla \times \mathbf{E}_1 + \nabla \times \mathbf{E}_2 = -\frac{\partial \mathbf{B}_1}{\partial t} - \frac{\partial \mathbf{B}_2}{\partial t}$$
$$\nabla \times \mathbf{B}_1 + \nabla \times \mathbf{B}_2 = \mu_0\,\mathbf{j}_1 + \mu_0\,\mathbf{j}_2 + \epsilon_0\mu_0\frac{\partial \mathbf{E}_1}{\partial t} + \epsilon_0\mu_0\frac{\partial \mathbf{E}_2}{\partial t}$$

which, if we combine like terms, simplifies to

$$\nabla \cdot (\mathbf{E}_1 + \mathbf{E}_2) = \frac{1}{\epsilon_0}(\rho_1 + \rho_2)$$
$$\nabla \cdot (\mathbf{B}_1 + \mathbf{B}_2) = 0$$
$$\nabla \times (\mathbf{E}_1 + \mathbf{E}_2) = -\frac{\partial (\mathbf{B}_1 + \mathbf{B}_2)}{\partial t}$$

$$\nabla \times (\mathbf{B}_1 + \mathbf{B}_2) = \mu_0 (\mathbf{j}_1 + \mathbf{j}_2) + \epsilon_0 \mu_0 \frac{\partial (\mathbf{E}_1 + \mathbf{E}_2)}{\partial t}$$

Since the net charge and current distributions are of course additive (so that $\rho_{net} = \rho_1 + \rho_2$ and $\mathbf{j}_{net} = \mathbf{j}_1 + \mathbf{j}_2$), we have

$$\begin{aligned}
\nabla \cdot (\mathbf{E}_1 + \mathbf{E}_2) &= \frac{1}{\epsilon_0} \rho_{net} \\
\nabla \cdot (\mathbf{B}_1 + \mathbf{B}_2) &= 0 \\
\nabla \times (\mathbf{E}_1 + \mathbf{E}_2) &= -\frac{\partial (\mathbf{B}_1 + \mathbf{B}_2)}{\partial t} \\
\nabla \times (\mathbf{B}_1 + \mathbf{B}_2) &= \mu_0 \mathbf{j}_{net} + \epsilon_0 \mu_0 \frac{\partial (\mathbf{E}_1 + \mathbf{E}_2)}{\partial t}
\end{aligned}$$

If we compare these relations to the Maxwell equations we would have for the net electric and magnetic fields of the combined distribution of charge and current,

$$\begin{aligned}
\nabla \cdot \mathbf{E}_{net} &= \frac{1}{\epsilon_0} \rho_{net} \\
\nabla \cdot \mathbf{B}_{net} &= 0 \\
\nabla \times \mathbf{E}_{net} &= -\frac{\partial \mathbf{B}_{net}}{\partial t} \\
\nabla \times \mathbf{B}_{net} &= \mu_0 \mathbf{j}_{net} + \epsilon_0 \mu_0 \frac{\partial \mathbf{E}_{net}}{\partial t}
\end{aligned}$$

we see that the net electric and magnetic fields are given by

$$\mathbf{E}_{net} = \mathbf{E}_1 + \mathbf{E}_2 \qquad\qquad \mathbf{B}_{net} = \mathbf{B}_1 + \mathbf{B}_2$$

This additivity is known as superposition or, if you have a bit more wind in you, as the principle of superposition. This basic property of electromagnetism is not only very important in principle, but as a practical matter makes the calculation of electric and magnetic fields much easier: to get the electric field due, for example, to an aggregation of point charges, you need merely take the vector sum of the electric fields to which each of the charges would individually give rise.

14.8 The Potential Functions

As was shown in §2.7.1, the magnetic version of Gauss's law,

$$\nabla \cdot \mathbf{B} = 0 \tag{14.1b}$$

implies that the magnetic field $\mathbf{B}$ is a pure curl and that it is therefore possible to find some vector field $\mathbf{A}$ such that

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{14.18}$$

This vector field $\mathbf{A}$ is known as the vector potential, and eq. (14.18) reduces the magnetic Gauss's law to an identity:

$$\begin{aligned}
\nabla \cdot \mathbf{B} &= \nabla \cdot (\nabla \times \mathbf{A}) \\
&= \left(\hat{\mathbf{x}} \frac{\partial}{\partial x} + \hat{\mathbf{y}} \frac{\partial}{\partial y} + \hat{\mathbf{z}} \frac{\partial}{\partial z} \right) \cdot \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ A_x & A_y & A_z \end{vmatrix} \\
&= \cdots \\
&= 0
\end{aligned}$$

where with the ellipsis we have omitted a lot of tedious but straightforward calculation that you can easily work out for yourself simply by doing out the components.

More generally, it was shown in §2.7.3 that any vector field can always be decomposed into two parts, a pure gradient and a pure curl, so that in particular it must be possible to write the electric field $\mathbf{E}$ in the form

$$\mathbf{E} = -\nabla\varphi + \nabla \times \mathbf{P}$$

where φ is a scalar field (called the scalar potential) and $\mathbf{P}$ is a vector field.[17] If we use this and our expression (14.18) for the magnetic field in Faraday's law (14.1c), we have

$$\begin{aligned}
\nabla \times \mathbf{E} &= -\frac{\partial \mathbf{B}}{\partial t} \\
\nabla \times (-\nabla\varphi + \nabla \times \mathbf{P}) &= -\frac{\partial}{\partial t} (\nabla \times \mathbf{A}) \\
\nabla \times \nabla\varphi + \nabla \times (\nabla \times \mathbf{P}) &= -\frac{\partial}{\partial t} (\nabla \times \mathbf{A}) \\
\nabla \times (\nabla \times \mathbf{P}) &= -\nabla \times \frac{\partial \mathbf{A}}{\partial t}
\end{aligned}$$

where we have noted, as you can again easily verify by simply doing it out in terms of components, that $\nabla \times \nabla\varphi = 0$.[18] This leads us to conclude that

[17] We have included a negative sign on the gradient term so that the scalar potential φ will match up with the scalar potential conventionally used in electromagnetism.

[18]
$$\nabla \times \nabla\varphi = \begin{vmatrix} \hat{\mathbf{x}} & \hat{\mathbf{y}} & \hat{\mathbf{z}} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ \frac{\partial\varphi}{\partial x} & \frac{\partial\varphi}{\partial y} & \frac{\partial\varphi}{\partial z} \end{vmatrix} = \cdots = 0$$

the $\nabla \times \mathbf{P}$ part of $\mathbf{E}$ is $-\partial\mathbf{A}/\partial t$:

$$\mathbf{E} = -\nabla\varphi - \frac{\partial \mathbf{A}}{\partial t} \tag{14.19}$$

If we use the expressions (14.19) and (14.18) for the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$ in terms of the scalar potential φ and vector potential $\mathbf{A}$, then, as we just saw, the magnetic Gauss's law (14.1a) and Faraday's law (14.1c) become identities, so that eqq. (14.19) and (14.18) can replace these two Maxwell equations. We may therefore equivalently write the four Maxwell equations (14.1) in the form

$$\mathbf{E} = -\nabla\varphi - \frac{\partial \mathbf{A}}{\partial t} \tag{14.20a}$$

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{14.20b}$$

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\rho \tag{14.20c}$$

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} + \epsilon_0\mu_0 \frac{\partial \mathbf{E}}{\partial t} \tag{14.20d}$$

At the moment expressing $\mathbf{E}$ and $\mathbf{B}$ in terms of potential functions probably seems like an unnecessary complication, but these relations will turn out to be very useful later on.

14.8.1 Gauge Transforms & Gauge Symmetry

While we're at it, we might as well note that there is an abstract symmetry in eqq. (14.20a) and (14.20b) known as a gauge symmetry: if, for some arbitrary function Λ, we make the shifts

$$\mathbf{A} \to \mathbf{A} + \nabla\Lambda \tag{14.21a}$$

$$\varphi \to \varphi - \frac{\partial \Lambda}{\partial t} \tag{14.21b}$$

in the values of φ and $\mathbf{A}$, then eqq. (14.20a) and (14.20b), and consequently all the physics arising from them, are unaffected:

$$\begin{aligned}
\mathbf{E} &= -\nabla\varphi - \frac{\partial \mathbf{A}}{\partial t} \\
&\to -\nabla\left(\varphi - \frac{\partial \Lambda}{\partial t}\right) - \frac{\partial}{\partial t}(\mathbf{A} + \nabla\Lambda) \\
&= -\nabla\varphi - \frac{\partial \mathbf{A}}{\partial t} + \left(\nabla\frac{\partial \Lambda}{\partial t} - \frac{\partial}{\partial t}\nabla\Lambda\right) \\
&= -\nabla\varphi - \frac{\partial \mathbf{A}}{\partial t}
\end{aligned}$$

$$\begin{aligned}\mathbf{B} &= \nabla \times \mathbf{A} \\ &\rightarrow \nabla \times (\mathbf{A} + \nabla\Lambda) \\ &= \nabla \times \mathbf{A} + \nabla \times \nabla\Lambda \\ &= \nabla \times \mathbf{A}\end{aligned}$$

where we have noted that the order of the derivatives of Λ does not matter and that, as you can verify simply by doing out the components, $\nabla \times \nabla$ acting on any scalar function vanishes identically.

The sort of shift made in the values of φ and $\mathbf{A}$ in eqq. (14.21) is known as a gauge transform and is of immense importance, far beyond its practical usefulness for simplifying certain calculations: symmetry turns out to be *the* fundamental principle in physics — the physics of the universe is determined by its symmetries — , and, as we will show in §22.6.1, it is the gauge symmetry associated with the gauge transform (14.21) that gives rise to electromagnetism and the Maxwell equations. Geometrically, this gauge symmetry follows from the symmetry of the unit circle: associated with every point in spacetime is a circle, and as a direct consequence of the rotational symmetry of this circle, there must be electromagnetic fields and interactions of exactly the sort specified by the Maxwell equations. In higher-dimensional theories such as string theory, these circles constitute one of the curled-up extra dimensions you may have read about.[19]

14.9 Light, Locality, & Relativity

One very important property of the Maxwell equations (14.1) and the force relations

$$\mathbf{F}_{electric} = q\mathbf{E} \qquad\qquad \mathbf{F}_{magnetic} = q\mathbf{v} \times \mathbf{B}$$

is that they are local — there is no action at a distance: the divergence of the electric field at a given point in spacetime is determined by the charge density *only at that same spacetime point*, and likewise the curl of the magnetic field at a given point in spacetime is determined by the current density *only at that same spacetime point*. Similarly, in the force relations the electric and magnetic forces experienced by a charge are determined solely by the values of the electric and magnetic fields *at the location of the charge*.

If we had more time and math at our disposal, we could show that changes in the electric and magnetic fields due to movement of charge propagate outward from their sources at the speed of light; the electric and magnetic field values at any given point in spacetime depend on the configuration of

[19] Or, if you haven't read about them, they are discussed briefly in §22.7.

charge and current on that point's backward light cone. Though we cannot establish this in full generality, we can show that the Maxwell equations can be combined to yield a wave equation, solutions to which propagate at the speed of light. Taking the curl of both sides of Faraday's law (14.1c), we have

$$\nabla \times (\nabla \times \mathbf{E}) = \nabla \times \left(-\frac{\partial \mathbf{B}}{\partial t} \right) \tag{14.22}$$

If we use the general vector-calculus result [20]

$$\nabla \times (\nabla \times \mathbf{E}) = \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E}$$

and Gauss's and Ampère's laws

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\,\rho \tag{14.1a}$$

$$\nabla \times \mathbf{B} = \mu_0\,\mathbf{j} + \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} \tag{14.1d}$$

eq. (14.22) becomes

$$\begin{aligned} \nabla(\nabla \cdot \mathbf{E}) - \nabla^2 \mathbf{E} &= \nabla \times \left(-\frac{\partial \mathbf{B}}{\partial t} \right) \\ \nabla\left(\frac{1}{\epsilon_0}\,\rho \right) - \nabla^2 \mathbf{E} &= -\frac{\partial}{\partial t}(\nabla \times \mathbf{B}) \\ &= -\frac{\partial}{\partial t}\left(\mu_0\,\mathbf{j} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) \end{aligned}$$

If we are in a region where there is no net charge and no movement of charge (such as a vacuum), then $\rho = 0$ and $\mathbf{j} = 0$, so that this reduces to

$$-\nabla^2 \mathbf{E} = -\mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2}$$

or

$$\nabla^2 \mathbf{E} = \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{14.23}$$

[20] To verify this relation, you need merely have the patience to work out the components of

$$\left(\widehat{\mathbf{x}}\,\frac{\partial}{\partial x} + \widehat{\mathbf{y}}\,\frac{\partial}{\partial y} + \widehat{\mathbf{z}}\,\frac{\partial}{\partial z} \right) \times \left(\left(\widehat{\mathbf{x}}\,\frac{\partial}{\partial x} + \widehat{\mathbf{y}}\,\frac{\partial}{\partial y} + \widehat{\mathbf{z}}\,\frac{\partial}{\partial z} \right) \times (E_x\,\widehat{\mathbf{x}} + E_y\,\widehat{\mathbf{y}} + E_z\,\widehat{\mathbf{z}}) \right)$$

on the left-hand side, and

$$\left(\widehat{\mathbf{x}}\,\frac{\partial}{\partial x} + \widehat{\mathbf{y}}\,\frac{\partial}{\partial y} + \widehat{\mathbf{z}}\,\frac{\partial}{\partial z} \right) \left(\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} \right) - \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \right) (E_x\,\widehat{\mathbf{x}} + E_y\,\widehat{\mathbf{y}} + E_z\,\widehat{\mathbf{z}})$$

on the right-hand side.

Similarly, taking the curl of both sides of Ampère's law (14.1d), we have

$$\nabla \times (\nabla \times \mathbf{B}) = \nabla \times \left(\mu_0 \mathbf{j} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right)$$

which, by means of

$$\nabla \times (\nabla \times \mathbf{B}) = \nabla(\nabla \cdot \mathbf{B}) - \nabla^2 \mathbf{B}$$

and the magnetic Gauss's and Faraday's laws

$$\nabla \cdot \mathbf{B} = 0 \tag{14.1b}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{14.1c}$$

becomes

$$\begin{aligned} \nabla(\nabla \cdot \mathbf{B}) - \nabla^2 \mathbf{B} &= \nabla \times \left(\mu_0 \mathbf{j} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \right) \\ -\nabla^2 \mathbf{B} &= \mu_0 \nabla \times \mathbf{j} + \mu_0 \epsilon_0 \frac{\partial}{\partial t} (\nabla \times \mathbf{E}) \\ &= \mu_0 \nabla \times \mathbf{j} + \mu_0 \epsilon_0 \frac{\partial}{\partial t} \left(-\frac{\partial \mathbf{B}}{\partial t} \right) \\ &= \mu_0 \nabla \times \mathbf{j} - \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{B}}{\partial t^2} \end{aligned}$$

In a vacuum, where $\rho = 0$ and $\mathbf{j} = 0$, this reduces to

$$-\nabla^2 \mathbf{B} = -\mu_0 \epsilon_0 \frac{\partial^2 \mathbf{B}}{\partial t^2}$$

or

$$\nabla^2 \mathbf{B} = \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{B}}{\partial t^2} \tag{14.24}$$

Together, eqq. (14.23) and (14.24) constitute the wave equations corresponding to an electromagnetic wave. To see this, let's simplify the above relation down to one dimension: along an x axis, eq. (14.23) reduces to

$$\frac{\partial^2 \mathbf{E}}{\partial x^2} = \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2}$$

The solution to this equation is a harmonic traveling wave of the form

$$\mathbf{E} = \mathbf{E}_0 \sin(kx - \omega t) \tag{14.25}$$

where $\mathbf{E}_0$ is the amplitude of the wave, $k = 2\pi/\lambda$ is its wave number,[21] and ω is its angular frequency. We can verify that this is indeed a solution by substituting (14.25) into eq. (14.23):

$$\frac{\partial^2 \mathbf{E}}{\partial x^2} = \mu_0 \epsilon_0 \frac{\partial^2 \mathbf{E}}{\partial t^2}$$

$$\frac{\partial^2}{\partial x^2}\Big(\mathbf{E}_0 \sin(kx - \omega t)\Big) = \mu_0 \epsilon_0 \frac{\partial^2}{\partial t^2}\Big(\mathbf{E}_0 \sin(kx - \omega t)\Big)$$

$$-k^2 \mathbf{E}_0 \sin(kx - \omega t) = -\mu_0 \epsilon_0 \, \omega^2 \mathbf{E}_0 \sin(kx - \omega t)$$

The proposed wave solution (14.25) will therefore satisfy eq. (14.23) as long as we take

$$k^2 = \mu_0 \epsilon_0 \, \omega^2$$

or, in other words,

$$\frac{\omega}{k} = \frac{1}{\sqrt{\mu_0 \epsilon_0}} \tag{14.26}$$

If we recall that the period T of a wave oscillation is related to its frequency f by $T = 1/f$ and use $k = 2\pi/\lambda$ and $\omega = 2\pi f$, we have

$$\frac{\omega}{k} = \frac{2\pi f}{2\pi/\lambda} = \lambda f = \frac{\lambda}{T}$$

Now, λ/T, the wavelength over the period, is just the distance the wave travels forward each cycle divided by the time for each cycle — that is, the velocity with which the wave is propagating. Our condition (14.26) is therefore telling us that the speed at which our electromagnetic wave is propagating is

$$\frac{1}{\sqrt{\mu_0 \epsilon_0}}$$

If you plug in the values of ϵ_0 and μ_0, the numerical result will in fact match the speed of light, c. The Maxwell equations thus predict not only the existence of light, but its speed.

Relativity is also already built into Maxwell equations, and in fact the above prediction of the value of the speed of light in terms of basic electric and magnetic constants, without any reference to a particular observer's frame, is what later led Einstein to postulate that the speed of light is the same in all reference frames and to formulate the special theory of relativity. And if you go back and apply special relativity to the Maxwell equations, you can see that electric and magnetic fields are really the same field; they are just two different manifestations of a single electromagnetic field. Though a full proof is once again beyond our scope, we can illustrate the general effect by

[21] You can take $k = 2\pi/\lambda$ as the definition of wave number in terms of the wavelength λ.

working out what happens for a simple configuration of charge and current if we borrow a couple of results for electric and magnetic fields from later chapters and apply time dilation and length contraction to see what these fields look like to two observers in relative motion.

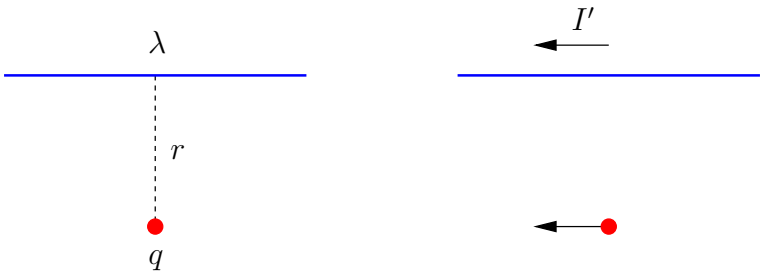


Figure 14.2: A Line and a Point Charge in Two Reference Frames

Suppose we have a positive point charge q a perpendicular distance r from an infinite line of positive charge λ per unit length, both at rest in our reference frame (the unprimed frame), as shown on the left side of fig. (14.2). From the perspective of an observer moving to the right at speed v (the primed frame), the line and point charges are both moving to the left at speed v, as shown on the right side of fig. (14.2); in particular, from this observer's perspective the motion of the line of charge constitutes a current that we will call I'. Our first task will be to determine how the forces according to us are related to the forces according to the moving observer, at least along the direction perpendicular to the relative motion. To accomplish this, we write these perpendicular force components in our frame in the form

$$F_\perp = \frac{dp_\perp}{dt}$$

Because there are no spatial relativistic effects along directions perpendicular to the relative motion of the two frames, the moving observer will agree with our value for the perpendicular component of the momentum, so that $dp'_\perp = dp_\perp$. There will, however, be disagreement about the time interval: to the moving observer, we are in motion and therefore experiencing time dilation; from the perspective of the moving observer, our clocks are running slow by a factor of $\gamma = 1/\sqrt{1 - v^2/c^2} = 1/\sqrt{1 - \beta^2}$, so that what is a time interval dt to us will, according to the moving observer, be of duration

$$dt' = \gamma\, dt$$

Thus the perpendicular force $F'_\perp$ according to the moving observer is related to the perpendicular force $F_\perp$ according to us by

$$F'_\perp = \frac{dp'_\perp}{dt'} = \frac{dp_\perp}{\gamma\, dt} = \frac{1}{\gamma}\, F_\perp \tag{14.27}$$

Next we ask specifically how the electric force exerted by the line charge on the point charge according to us is related to the electric force exerted by the line charge on the point charge according to the moving observer. In the reference frame of the moving observer, the line charge is moving and therefore experiences length contraction, which would increase the charge λ per unit length on it, and consequently both the electric field of the line and the electric force it exerts on the point charge, by a factor of γ:

$$\lambda' = \gamma\lambda \tag{14.28a}$$

$$F'_{electric} = \gamma F_{electric} \tag{14.28b}$$

In our frame, where there is no movement of charge, the electric force constitutes the entirety of the perpendicular force, so that

$$F_\perp = F_{electric} \tag{14.29}$$

There is thus a discrepancy between the perpendicular force and the electric force seen by the moving observer: putting together eqq. (14.27), (14.28b), and (14.29), we have

$$\begin{aligned} F'_\perp - F'_{electric} &= \frac{1}{\gamma} F_\perp - \gamma F_{electric} \\ &= \frac{1}{\gamma} F_{electric} - \gamma F_{electric} \\ &= \left(\frac{1}{\gamma^2} - 1\right) \gamma F_{electric} \\ &= \left(\frac{1}{\left(1/\sqrt{1-\beta^2}\right)^2} - 1\right) \gamma F_{electric} \\ &= -\beta^2 \gamma F_{electric} \end{aligned} \tag{14.30}$$

where the negative sign indicates that this difference in force is toward the line charge.

Our final task is to show that this discrepancy is accounted for by magnetic effects, a task that will require what may, from your present perspective, seem like gerrymandering of epic proportions.[22] First, as we will see in Chapter 15, the expression for the electric force in our frame is

$$F_{electric} = \frac{1}{4\pi\epsilon_0} \frac{2q\lambda}{r}$$

[22] Actually, as far as we know, Homer never used the term *gerrymander*.

If we use this together with $\lambda' = \gamma\lambda$ from eq. (14.28a) and $\beta = v/c$, and rearrange some factors, our discrepancy (14.30) can be re-expressed as

$$-\beta^2\gamma F_{electric} = -\left(\frac{v}{c}\right)^2 \gamma \frac{1}{4\pi\epsilon_0}\frac{2q\lambda}{r} = -\frac{1}{\epsilon_0 c^2}\, qv\, \frac{\gamma\lambda v}{2\pi r} = -\frac{1}{\epsilon_0 c^2}\, qv\, \frac{\lambda' v}{2\pi r}$$

And if, with almost divine foresight, we now artificially multiply and divide this by μ_0, and then use $c = 1/\sqrt{\epsilon_0\mu_0}$ from (14.26), this discrepancy can be further morphed into

$$-\frac{1}{\epsilon_0\mu_0\, c^2}\, qv\, \frac{\mu_0\lambda' v}{2\pi r} = -qv\, \frac{\mu_0\lambda' v}{2\pi r} \tag{14.31}$$

Consider now the current I' that the moving observer sees traveling down the line: if, as it moves to the left at speed v, the charge on the line undergoes a displacement ds' in time dt' according to the moving observer, then, by eq. (14.4), the current I' will be given by

$$I' = \frac{dq'}{dt'} = \frac{dq'}{ds'}\frac{ds'}{dt'} = \lambda' v$$

Our discrepancy (14.31) can therefore be rewritten

$$-qv\, \frac{\mu_0 I'}{2\pi r} \tag{14.32}$$

As we will see in Chapter 17, the magnetic field of an infinite line of current flowing to the left is, at the location of the point charge q,

$$\mathbf{B}' = \frac{\mu_0 I'}{2\pi r}\, \odot$$

Since the velocity of the charge q according to the moving observer is to the left, the corresponding magnetic force works out to

$$\begin{aligned} \mathbf{F}'_{magnetic} &= q\mathbf{v}' \times \mathbf{B}' \\ &= q(v\ \leftarrow) \times \left(\frac{\mu_0 I'}{2\pi r}\, \odot\right) \\ &= qv\, \frac{\mu_0 I'}{2\pi r}\ \uparrow \end{aligned}$$

Up the page being toward the line of charge, this exactly accounts for both the magnitude and direction of the discrepancy (14.32).

The moral of all of this is that, since what is moving depends on who you ask, effects that are electric to one observer will be partly magnetic to another (and, as it turns out, vice versa). To the observer in the reference frame where the line and point charge are at rest, there are only electric fields and only

electric forces between charges. But this same configuration of charge is in motion according to the moving observer, and in that observer's reference frame there are also magnetic fields and forces. Just as spatial displacements and time intervals transform into each other as you go from one reference frame to another, electric and magnetic fields transform into each other. And so, just as space and time become in relativity inseparable parts of a four-dimensional spacetime, electric and magnetic fields in relativity meld into single electromagnetic field.[23]

[23]Unlike spacetime, this electromagnetic field is not, however, as simple as a four-vector that transforms under the Lorentz transform (10.29): while a time and three spatial coordinates fit nicely into a four-component vector, the electric and magnetic fields each have three components — a total of six. As we will see explicitly in §22.6.1, it turns out that the electromagnetic field is actually a tensor, a 4×4 matrix $F^{\mu\nu}$, each index (that is, each row and column) of which transforms according to the Lorentz transform (10.29). This matrix turns out to be antisymmetric, so that $F^{\mu\nu} = -F^{\nu\mu}$. Of the $4 \times 4 = 16$ elements of this matrix, the four diagonal elements all vanish because $F^{\mu\mu} = -F^{\mu\mu}$ requires $F^{\mu\mu} = 0$. And of the 12 remaining nonzero elements, $F^{\mu\nu} = -F^{\nu\mu}$ means that only six are independent — exactly the number needed to contain both the electric and magnetic fields. Just in case you were curious.

Chapter 15

Electrostatics

In electrostatics we deal with static electric fields, that is, fields that have no time dependence. Otherwise, *electrostatics* would be something of a misnomer, wouldn't it? Anyway, the configurations of charge that give rise to these fields must therefore have been static for long enough that changes in the field, which propagate outward from their sources at the speed of light, will have had time to reach the locations with which we are concerned.

15.1 Applications of Gauss's Law

In addition to its importance in principle, in practice Gauss's law

$$\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} = \frac{1}{\epsilon_0}\,q_{enclosed} \tag{14.2a}$$

enables us to determine the electric fields of highly symmetric distributions of charge.[1] The procedure is very straightforward:

- First surmise the direction and dependence of the **E** field from the symmetry of the distribution.
- Carry out the surface integration on the left-hand side over a Gaussian surface S that reflects the distribution's symmetry.
- Determine how much charge $q_{enclosed}$ is interior to that Gaussian surface.

And from there it's child's play to solve for **E**.

[1] While we're on matters of principle and practice, we should note that Gauss's law could in principle be used to obtain a result for the electric field of any distribution of charge; it is just that in the absence of symmetry the surface integration is, as a practical matter, intractable.

15.1.1 Spherical Charge Distributions

Suppose we have a total charge q uniformly (that is, evenly) distributed over a spherical shell of radius a. Spherical coordinates are most natural to the spherical symmetry of this distribution of charge, and in spherical coordinates the most general expression for the electric field is

$$\mathbf{E}(\mathbf{r}) = E_r(r,\theta,\varphi)\,\hat{\mathbf{r}} + E_\theta(r,\theta,\varphi)\,\widehat{\boldsymbol{\theta}} + E_\varphi(r,\theta,\varphi)\,\widehat{\boldsymbol{\varphi}}$$

Now, because the distribution of charge is spherically symmetric, we can rotate it by any angle in the θ or φ directions or flip it about any plane through the origin and it will look the same after the rotation or flip as it did before it. Since the electric field must have the same symmetry as the distribution of charge that gives rise to it, $\mathbf{E}$ cannot have any $\widehat{\boldsymbol{\theta}}$ or $\widehat{\boldsymbol{\varphi}}$ components; if it did, then such rotations or flips would change the direction of $\mathbf{E}$, which would violate the symmetry. $\mathbf{E}$ therefore reduces to

$$\mathbf{E}(\mathbf{r}) = E_r(r,\theta,\varphi)\,\hat{\mathbf{r}}$$

Next we note that $\mathbf{E}$ also cannot have any dependence on θ or φ; if it did, then the rotations or flips we just mentioned would change the value of $\mathbf{E}$, which would again violate the symmetry of the charge distribution. So $\mathbf{E}$ further reduces to

$$\mathbf{E}(\mathbf{r}) = E_r(r)\,\hat{\mathbf{r}} = E(r)\,\hat{\mathbf{r}}$$

That is, the electric field must be in the radial direction, and its magnitude can depend only on our distance from the center of the shell.

The Gaussian surfaces that reflect the symmetry of the charge distribution are spheres concentric with the shell of charge. Over such a Gaussian surface S, the outward normal is simply $\hat{\mathbf{r}}$, and the left-hand side of Gauss's law works out to

$$\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} = \oint_S dA\,\hat{\mathbf{r}}\cdot\Big(E(r)\,\hat{\mathbf{r}}\Big) = \oint_S dA\,E(r) = E(r)\oint_S dA = E(r)\,4\pi r^2$$

where we have noted that since r and hence $E(r)$ are constant over the spherical Gaussian surface, we can pull $E(r)$ outside of the integration, which then just gives us the total area $4\pi r^2$ of the Gaussian surface.

On the right-hand side of Gauss's law, how much charge is enclosed by the Gaussian surface depends on its radius: if the radius r of the Gaussian surface is greater than the radius a of the shell of charge, then all of the charge q on the shell is enclosed; if the radius r of the Gaussian surface is less than the radius a of the shell of charge, then none of the charge q on the shell is enclosed. Thus Gauss's law

$$\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} = \frac{1}{\epsilon_0}\,q_{enclosed}$$

gives us

$$E(r)\,4\pi r^2 = \frac{1}{\epsilon_0}\begin{cases} q & (r > a) \\ \\ 0 & (r < a) \end{cases}$$

and hence

$$E(r) = \begin{cases} \dfrac{1}{4\pi\epsilon_0}\,\dfrac{q}{r^2} & (r > a) \\ \\ 0 & (r < a) \end{cases} \tag{15.1}$$

as our result for the electric field of a uniform shell of radius a and charge q.

Note that outside the shell the field falls off as $1/r^2$: in order to keep the total electric flux constant, the field must compensate for the r^2 growth in the area of the spherical Gaussian surfaces by falling off as $1/r^2$. Also note that the direction of the field depends on the sign on q: for a positive charge q, the field is radially outward, away from the charge; for negative charge q, the field is inward, toward the charge. This is an example of the general result that

- The electric field of a positive charge points away from it.
- The electric field of a negative charge points toward it.

A special case of (15.1) of great importance is a point charge, which can be considered a spherical shell of zero radius, and for which we therefore have

$$E = \frac{1}{4\pi\epsilon_0}\,\frac{q}{r^2} \tag{15.2}$$

So from the outside, uniform spherical shells of charge look, electrically, just like point charges. In fact, since it can be regarded as an amalgamation of uniform shells, any rotationally symmetric solid sphere or spherical shell of finite thickness will, at points exterior to it, be electrically equivalent to a point charge.

The simplest such spherically symmetric distribution of charge is a solid sphere of radius a with a total charge q uniformly spread throughout its interior. The symmetry of this distribution, and hence the symmetry of $\mathbf{E}$ and our result for the left-hand side of Gauss's law, will be the same as for the case of the spherical shell. The difference will be on the right-hand side of Gauss's law: while we again enclose the full charge q of the sphere when $r > a$, when $r < a$ we pick up only that fraction of the sphere's total charge that lies within the radius r of our Gaussian surface. And since the charge density within the sphere is uniform, that fraction is proportional to volume:

$$q_{enclosed} = \frac{\frac{4}{3}\pi r^3}{\frac{4}{3}\pi a^3}\,q = \left(\frac{r}{a}\right)^3 q$$

Thus for the solid sphere of charge Gauss's law gives

$$E(r)\,4\pi r^2 = \frac{1}{\epsilon_0}\begin{cases} q & (r > a) \\ \left(\dfrac{r}{a}\right)^3 q & (r < a) \end{cases}$$

and hence

$$E(r) = \begin{cases} \dfrac{1}{4\pi\epsilon_0}\,\dfrac{q}{r^2} & (r > a) \\ \dfrac{1}{4\pi\epsilon_0}\,\dfrac{qr}{a^3} & (r < a) \end{cases} \tag{15.3}$$

Note that as $r \to 0$, $E \sim r^1 \to 0$: although the Gaussian surface area is shrinking as r^2, the charge enclosed is shrinking even faster, as r^3.

What if the charge distribution within the sphere is a function of r, so that, while still spherically symmetric, it is no longer uniform? In this case, $q_{enclosed}$ is no longer proportional to the volume enclosed; we have to do a volume integration. Suppose, for example, that the charge density ρ within the sphere goes as $1/r$, that is, is of the form

$$\rho = C\frac{1}{r}$$

where C is a constant. If the total charge on the sphere is q, then, using $\rho = dq/dV$ from eq. (14.3) in the form $dq = \rho\,dV$, we have

$$\begin{aligned} q &= \int_{\text{sphere}} dq \\ &= \int_{\text{sphere}} dV\,\rho \\ &= \int_{\text{sphere}} r^2 \sin\theta\,dr\,d\theta\,d\varphi\,C\frac{1}{r} \\ &= C\int_0^a r\,dr\int_0^\pi \sin\theta\,d\theta\int_0^{2\pi} d\varphi \\ &= C\,(\tfrac{1}{2}a^2)\,(2)\,(2\pi) \\ &= C\,2\pi a^2 \end{aligned}$$

which yields

$$C = \frac{q}{2\pi a^2}$$

and hence

$$\rho = C\frac{1}{r} = \frac{q}{2\pi a^2}\frac{1}{r}$$

So the charge $q_{enclosed}$ enclosed by the volume V of a Gaussian surface of radius r would be, by a very similar integration,

$$\begin{aligned} q_{enclosed} &= \int_V dV\, \rho \\ &= \int_V r^2 \sin\theta\, dr\, d\theta\, d\varphi\, \frac{q}{2\pi a^2 r} \\ &= \frac{q}{2\pi a^2} \int_0^r r\, dr \int_0^\pi \sin\theta\, d\theta \int_0^{2\pi} d\varphi \\ &= \frac{q}{2\pi a^2} \left(\tfrac{1}{2} r^2\right) (2)\, (2\pi) \\ &= \left(\frac{r}{a}\right)^2 q \end{aligned}$$

And from here, the rest of the calculation of $\mathbf{E}$ would be the same as that for the uniform sphere of charge.

You should be able to apply Gauss's law to the following spherically symmetric distributions of charge:

- A point charge.
- A solid sphere, of uniform or nonuniform charge density.
- A spherical shell, of either infinitesimal or finite thickness.

Note that our results, such as

$$E = \frac{1}{4\pi\epsilon_0} \frac{q}{r^2}$$

for the field of a point charge, are valid only for stationary charges: while Gauss's law would still give

$$\oint_S dA\, \hat{\mathbf{n}} \cdot \mathbf{E} = \frac{1}{\epsilon_0}\, q$$

even if the charge q were moving inside the Gaussian surface S, any movement of the charge is necessarily in a particular direction, and this "special" direction of motion would break the spherical symmetry on which eq. (15.2) and its ilk depend. So while Gauss's law is valid even when charge is in motion, our results for electric fields in this section are not.

15.1.2 Cylindrical Charge Distributions

Now suppose we have a uniform linear charge density λ (charge per unit length) distributed over an infinite cylindrical shell of radius a. Cylindrical coordinates are most natural to the cylindrical symmetry of this distribution of charge, and in cylindrical coordinates the most general expression for the electric field is

$$\mathbf{E}(\mathbf{r}) = E_r(r,\theta,z)\,\hat{\mathbf{r}} + E_\theta(r,\theta,z)\,\widehat{\boldsymbol{\theta}} + E_\varphi(r,\theta,z)\,\hat{\mathbf{z}}$$

Because the distribution of charge is cylindrically symmetric, we can rotate it by any angle θ or flip it about the z axis and it will look the same after the rotation or flip as it did before it. Since the electric field must share this same symmetry, $\mathbf{E}$ cannot have any $\widehat{\boldsymbol{\theta}}$ or $\hat{\mathbf{z}}$ components; if it did, then such rotations or flips would change the direction of $\mathbf{E}$, which would violate the symmetry. $\mathbf{E}$ therefore reduces to

$$\mathbf{E}(\mathbf{r}) = E_r(r,\theta,z)\,\hat{\mathbf{r}}$$

Next we note that $\mathbf{E}$ also cannot have any dependence on θ; if it did, then the rotations or flips we just mentioned would change the value of $\mathbf{E}$, which would again violate the symmetry of the charge distribution. Similarly, $\mathbf{E}$ cannot have any dependence on z, because shifting up or down along the axis of an *infinite* cylinder cannot make any difference. So $\mathbf{E}$ further reduces to

$$\mathbf{E}(\mathbf{r}) = E_r(r)\,\hat{\mathbf{r}} = E(r)\,\hat{\mathbf{r}}$$

That is, the electric field must be in the radial direction, and its magnitude can depend only on our distance from the axis of the shell.

The Gaussian surfaces that reflect the symmetry of the charge distribution are cylinders coaxial with the shell of charge. In order to form a closed surface, our Gaussian cylinders must, however, be of a finite length, say ℓ, and we have to break up our integration over such a Gaussian surface S into an integration over the side of the cylinder and integrations over the end-caps:

$$\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} = \int_{\text{side}} dA\,\hat{\mathbf{n}}\cdot\mathbf{E} + \int_{\text{end-caps}} dA\,\hat{\mathbf{n}}\cdot\mathbf{E}$$

Over the end-caps, the outward normal to the Gaussian cylinder will be in the $\pm\hat{\mathbf{z}}$ direction, so that

$$\hat{\mathbf{n}}\cdot\mathbf{E} \sim \hat{\mathbf{z}}\cdot\hat{\mathbf{r}} = 0$$

So only the side of the Gaussian cylinder, over which the outward normal is $\hat{\mathbf{r}}$, contributes to the surface integration:

$$\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} = \int_{\text{side}} dA\,\hat{\mathbf{n}}\cdot\mathbf{E} = \int_{\text{side}} dA\,\hat{\mathbf{r}}\cdot\Big(E(r)\,\hat{\mathbf{r}}\Big) = E(r)\int_{\text{side}} dA = E(r)\,2\pi r\ell$$

where we have noted that since r and hence $E(r)$ are constant over the side of the Gaussian cylinder, we can pull $E(r)$ outside of the integration, which then just gives us the area $2\pi r\ell$ of the side of the cylinder.

On the right-hand side of Gauss's law, how much charge is enclosed by the Gaussian cylinder depends on its radius: if the radius r of the Gaussian cylinder is greater than the radius a of the shell of charge, then all of the charge $\lambda\ell$ on the intersected section of the shell is enclosed; if the radius r of the Gaussian cylinder is less than the radius a of the shell of charge, then none of the charge on the shell is enclosed. Thus Gauss's law

$$\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} = \frac{1}{\epsilon_0}\,q_{enclosed}$$

gives us

$$E(r)\,2\pi r\ell = \frac{1}{\epsilon_0}\begin{cases} \lambda\ell & (r > a) \\ \\ 0 & (r < a) \end{cases}$$

and hence [2]

$$E(r) = \begin{cases} \dfrac{1}{4\pi\epsilon_0}\,\dfrac{2\lambda}{r} & (r > a) \\ \\ 0 & (r < a) \end{cases} \tag{15.4}$$

Note that, in contrast to the $1/r^2$ drop in the field of a spherical charge, the field of a cylinder of charge falls off as $1/r$. This is what we would expect for a one-dimensionally infinite distribution of charge: were we to obtain the field of the cylinder (as we in fact will in §15.3.3 and problem #26) by integrating the contributions from all of the infinitesimal charges dq that make it up, each of these contributions would be like the $1/r^2$ field of a point charge, and when integrated over the infinite length of the cylinder would yield a net field that falls off as $1/r$.

A special case of (15.4) is a line charge, which can be considered a cylindrical shell of zero radius, and for which we therefore have

$$E = \frac{1}{4\pi\epsilon_0}\,\frac{2\lambda}{r} \tag{15.5}$$

Another cylindrically symmetric distribution of charge is a solid cylinder of radius a with a uniform linear charge density λ spread throughout its

[2] We have refrained from simplifying to

$$\frac{1}{2\pi\epsilon_0}\,\frac{\lambda}{r}$$

in order to keep a consistent overall factor of $1/4\pi\epsilon_0$ out front. In more sensible systems of units, $1/4\pi\epsilon_0$ is just 1. You do whatever you want with the 2 and the 4.

interior. The symmetry of this distribution, and hence the symmetry of $\mathbf{E}$ and our result for the left-hand side of Gauss's law, will be the same as for the case of the cylindrical shell. The difference will be on the right-hand side of Gauss's law: while we again enclose the full charge $\lambda\ell$ of the intersected section of the cylinder when $r > a$, when $r < a$ we pick up only that fraction of the charge $\lambda\ell$ that lies within the radius r of our Gaussian cylinder. And since the charge density within the cylinder is uniform, this fraction will be proportional to cross-sectional area:

$$q_{enclosed} = \frac{\pi r^2 \ell}{\pi a^2 \ell}\, q = \left(\frac{r}{a}\right)^2 q$$

Thus for the solid cylinder of charge Gauss's law gives

$$E(r)\, 2\pi r\ell = \frac{1}{\epsilon_0} \begin{cases} \lambda\ell & (r > a) \\ \left(\dfrac{r}{a}\right)^2 \lambda\ell & (r < a) \end{cases}$$

and hence

$$E(r) = \begin{cases} \dfrac{1}{4\pi\epsilon_0}\, \dfrac{2\lambda}{r} & (r > a) \\ \dfrac{1}{4\pi\epsilon_0}\, \dfrac{2\lambda r}{a^2} & (r < a) \end{cases} \tag{15.6}$$

You should be able to apply Gauss's law to the following spherically symmetric distributions of charge:

- An infinite line.
- An infinite solid cylinder, of either uniform or nonuniform charge density.
- An infinite cylindrical shell, of either infinitesimal or finite thickness.

15.1.3 Planar Charge Distributions

Finally, suppose we have a uniform surface charge density σ (charge per unit area) distributed over an infinite plane. Cartesian coordinates, with the sheet of charge in the xy plane, are most natural to the symmetry of this distribution of charge. In Cartesian coordinates the most general expression for the electric field is

$$\mathbf{E}(\mathbf{r}) = E_x(x, y, z)\, \hat{\mathbf{x}} + E_y(x, y, z)\, \hat{\mathbf{y}} + E_z(x, y, z)\, \hat{\mathbf{z}}$$

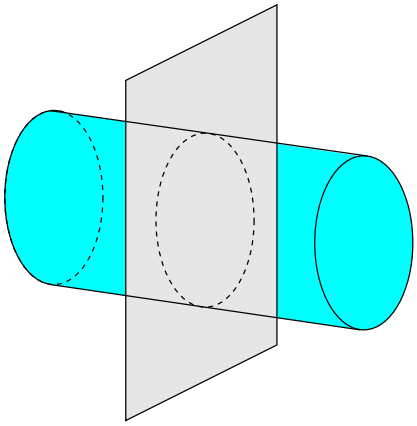


Figure 15.1: Cubist Representation of a Gaussian Pillbox

Since we can flip the sheet of charge about any axis and it will look the same after flip as it did before it, and since the electric field must share this same symmetry, $\mathbf{E}$ cannot have any $\hat{\mathbf{x}}$ or $\hat{\mathbf{y}}$ components; if it did, then such flips would reverse the direction of those components of $\mathbf{E}$, which would violate the symmetry. $\mathbf{E}$ therefore reduces to

$$\mathbf{E}(\mathbf{r}) = E_z(x, y, z)\,\hat{\mathbf{z}}$$

Next we note that $\mathbf{E}$ also cannot have any dependence on x or y, because shifting along an *infinite* sheet cannot make any difference. So $\mathbf{E}$ further reduces to

$$\mathbf{E}(\mathbf{r}) = E_z(z)\,\hat{\mathbf{z}}$$

Finally, we note that flipping the z axis should also make no difference, in the sense that a field pointed away from the sheet on one side should be pointed away from it on the other side as well, and likewise for a field pointed toward the sheet. Since flipping the z axis is equivalent to interchanging z and $-z$, this means that

$$E_z(-z) = -E_z(z)$$

where the negative sign on the right-hand side ensures that the field is in opposite directions on opposite sides of the sheet — either toward the sheet on both sides, or away from it on both sides.

As shown, in the mingled styles of M.C. Escher and Grandma Moses, in fig. (15.1), the Gaussian surfaces that reflect the symmetry of the charge distribution are "pillboxes" that intersect the sheet of charge, with sides perpendicular to the sheet and planar end-caps parallel to it and an equal distance from it (so that one end-cap is at z and the other at $-z$). As long as they are planar and parallel to the sheet, the shape of the end-caps doesn't matter — they could be square, circular, whatever turns you on.[3]

[3]They also don't have to be cyan.

We'll suppose they have area A. We need to break up the integration over the surface of our Gaussian pillbox into an integration over its side and integrations over the end-caps:

$$\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} = \int\limits_{\text{side}} dA\,\hat{\mathbf{n}}\cdot\mathbf{E} + \int\limits_{\text{end-caps}} dA\,\hat{\mathbf{n}}\cdot\mathbf{E}$$

Over the side, the outward normal to the Gaussian pillbox will always be in some direction parallel to the sheet, while $\mathbf{E}$ is perpendicular to the sheet, so that $\hat{\mathbf{n}}\cdot\mathbf{E} = 0$. Therefore only the end-caps of the Gaussian pillbox contribute to the surface integration. For the end-cap on the $+z$ side, the outward normal is $+\hat{\mathbf{z}}$; for the end-cap on the $-z$ side, $-\hat{\mathbf{z}}$. Thus we have

$$\begin{aligned}
\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} &= \int\limits_{+z\text{ end-cap}} dA\,\hat{\mathbf{z}}\cdot\Big(E_z(z)\,\hat{\mathbf{z}}\Big) + \int\limits_{-z\text{ end-cap}} dA\,(-\hat{\mathbf{z}})\cdot\Big(E_z(-z)\,\hat{\mathbf{z}}\Big) \\
&= E_z(z)\int\limits_{+z\text{ end-cap}} dA - \;E_z(-z)\int\limits_{-z\text{ end-cap}} dA \\
&= \Big(E_z(z) - E_z(-z)\Big)A \\
&= \Big(E_z(z) + E_z(z)\Big)A \\
&= 2AE_z(z)
\end{aligned}$$

where we have noted that since z and hence the $E_z(\pm z)$ are constant over each end-cap of the Gaussian pillbox, we can pull the $E_z(\pm z)$ outside of the integrations, each of which then just gives us the area A of the end-cap.

On the right-hand side of Gauss's law, the charge enclosed by the Gaussian pillbox is the charge σA on the intersected section of the sheet. Thus Gauss's law

$$\oint_S dA\,\hat{\mathbf{n}}\cdot\mathbf{E} = \frac{1}{\epsilon_0}\,q_{enclosed}$$

gives us

$$2AE_z(z) = \frac{1}{\epsilon_0}\,\sigma A$$

and hence [4]

$$E = \frac{1}{4\pi\epsilon_0}\,2\pi\sigma \tag{15.7}$$

Note that whereas the field of a spherical charge drops off as $1/r^2$ and that of a cylindrical charge as $1/r$, the field of the infinite sheet of charge is independent of our distance z from the sheet. This is what we would expect

[4] Again, we are keeping a consistent overall factor of $1/4\pi\epsilon_0$ out front. You do whatever you want.

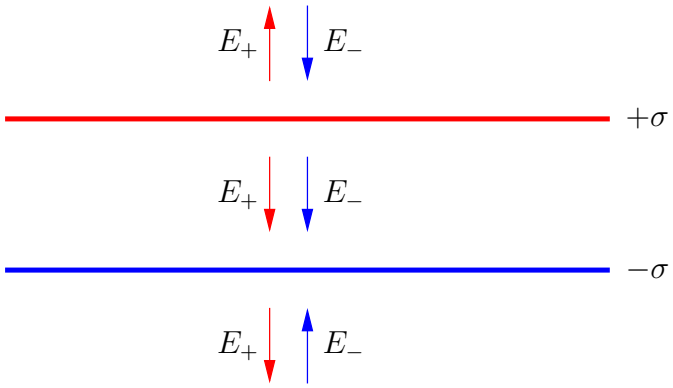


Figure 15.2: Electric Field of Parallel Plates

for a two-dimensionally infinite distribution of charge: were we to obtain the field of the sheet by integrating the contributions from all of the infinitesimal charges dq that make it up, each of these contributions would be like the $1/r^2$ field of a point charge, and when integrated over the infinite length and width of the sheet would yield a net field that falls off as r^{0}.[5]

You should be able to apply Gauss's law to

- Infinite sheets of charge.[6]

15.1.4 Superposition

Since, according to Gauss's law, the electric field **E** is linearly related to the charge distribution ρ, we can get results for a composite distribution of charge by superposition, that is, by simply adding up (as vectors, of course) the electric fields to which each of the various parts of the composite distribution individually gives rise.

Suppose, for example, that we have two infinite parallel sheets of charge, one of positive surface charge density $+\sigma$ and the other of equal negative surface charge density $-\sigma$, as depicted from a side view, rather abstractly, and for sheets that don't even pretend to be infinite, in fig. (15.2). By eq. (15.7), the contributions of each sheet to the net electric field are of magnitude $\sigma/2\epsilon_0$, in the directions shown in fig. (15.2): those of the positive sheet are everywhere directly away from the positive sheet, those of the negative sheet everywhere directly toward the negative sheet. This means that, when added

[5] A dependence like $\ln r$, which goes as r^0, therefore isn't ruled out. But we're only making a loose argument.

[6] How's that for an anticlimax?

as vectors, the fields of the two sheets will completely cancel each other everywhere outside the sheets, and in between the sheets[7] will reinforce each other to give a net field of σ/ϵ_0 that points from the positive toward the negative sheet.

Parallel plates like this are in fact an easy practical way to create a nice constant electric field within some region. You cannot, of course, make the plates infinite, but as long as they are large compared to the distance between them and you are not too close to an edge, the field will be approximately the constant field of infinite parallel plates.

15.2 Coulomb's Semibogus Law

Recall that the electric force on a point charge q is given by

$$\mathbf{F} = q\mathbf{E} \tag{14.7}$$

where $\mathbf{E}$ is the value of the electric field at the location of q. If we are dealing with two point charges, q_1 and q_2, separated by a distance r, then the electric field of q_2 at the location of q_1 will be

$$E = \frac{1}{4\pi\epsilon_0}\frac{q_2}{r^2}$$

so that the force on q_1 will be

$$\begin{aligned} F &= q_1 \frac{1}{4\pi\epsilon_0}\frac{q_2}{r^2} \\ &= \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r^2} \end{aligned} \tag{15.8}$$

Eq. (15.8) is known as Coulomb's semibogus law[8] and is the law for the force between two point charges. Or, to be more precise, almost the law for the force between two point charges. Recall that eq. (15.2) for the electric field of a point charge is valid only when the point charge giving rise to that field has been stationary long enough for its field to have propagated, at the speed of light, out to the points we are concerned with. For Coulomb's semibogus law to apply and be reciprocal, the two point charges in it must therefore have been at rest for at least as long as r/c.

Questions of bogosity aside, Gauss's law has now given us a huge bonus: in §15.1.1 we saw that at locations outside its radius a spherically symmetric distribution of charge is equivalent to a point charge, in the sense that it gives rise to the same electric field. And since it gives rise to the same electric field,

[7] Ha, ha!

[8] Actually, most people leave out the "semibogus."

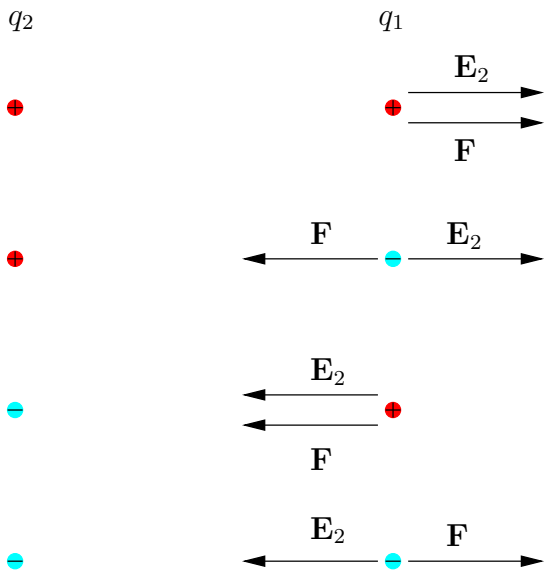


Figure 15.3: They Went That'a Way

it will also give rise to the same electric force. Now, Coulomb's semibogus law and Newton's law of gravity are both inverse-square laws,

$$F = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r^2} \qquad\qquad F = \frac{Gm_1 m_2}{r^2}$$

Mathematically, what holds for Coulomb's semibogus law must therefore hold for Newton's law of gravity as well, so that at locations outside their radii spherically symmetric distributions of mass must be equivalent gravitationally to point masses. And of course the results for the interior of uniform spherical shells of charge also carry over to uniform spherical shells of mass: just as the electric field and therefore the electric force vanish at radii interior to the shell of charge, the gravitational force exerted on masses inside of a shell of mass must vanish.

But back to Coulomb's semibogus law. If q_2 is a positive charge, its electric field $\mathbf{E}_2$ will be pointing away from it, so that the force

$$\mathbf{F} = q_1 \mathbf{E}_2$$

felt by q_1 will be in the same direction as $\mathbf{E}_2$ if q_1 is positive and in the direction opposite to $\mathbf{E}_2$ if q_1 is negative. That is, the force felt by q_1 will be away from q_2 when q_1 and q_2 are both positive and toward q_2 when q_1 is negative and q_2 is positive, as shown in the top two cases in fig. (15.3). And the same sort of results hold for the two cases when q_2 is negative, as shown in the bottom two cases in fig. (15.3). The upshot of all this is that $\mathbf{F} = q\mathbf{E}$ in conjunction with the general result that

- The electric field of a positive charge points away from it.
- The electric field of a negative charge points toward it.

is equivalent to the statement that

- Like charges repel.
- Unlike charges attract.

This rule for directions can be used to carry out the vector sums needed to get the net electric force on a point charge when there are electric forces exerted on it by more than one other charge. If continuous distributions of charge are involved, this vector sum will become an integration over the bits of charge $\rho\, dV$ (or $\sigma\, dA$ or $\lambda\, ds$) that make up the distribution — integrations very similar to those we will carry out in the next section.

15.3 Electric Fields by Direct Integration

Another technique for determining the electric field of a continuous distribution of charge is to directly integrate the contributions to the field from the infinitesimal bits of charge that make up the distribution: the contribution of each bit dq of charge will be like that of a point charge,

$$\frac{1}{4\pi\epsilon_0}\frac{dq}{r^2} \tag{15.9}$$

If the distribution is sufficiently symmetric that we can infer the direction of the net electric field, when we take the vector sum of the fields of the dq we need integrate only their components along that direction. In rather loose notation, the magnitude of the net field will then be given by

$$E_{net} = \frac{1}{4\pi\epsilon_0}\int \frac{dq}{r^2} \times (\text{trig factor})$$

where the "trig factor" extracts the component of the contributions (15.9) along the direction of the net field. This technique, illustrated by example in the following subsections, will allow us to determine the electric field of some distributions of charge that are not symmetric enough to be dealt with by Gauss's law.

15.3.1 Rings of Charge

Consider the case of a uniform thin ring of positive charge q and radius a. When we evaluate the electric field on the axis of the ring, at a distance

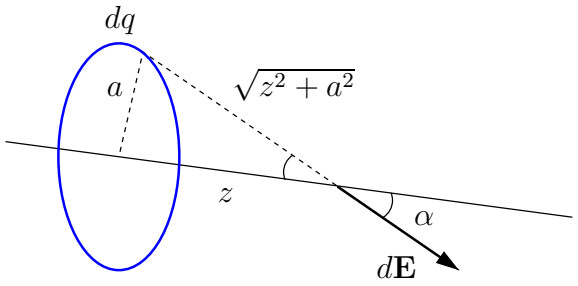


Figure 15.4: Field of a Ring of Charge

z from its center, our squared distance from each of the bits of charge dq around the ring is $a^2 + z^2$, as shown in fig. (15.4). The electric field of each bit dq is therefore of magnitude

$$dE = \frac{1}{4\pi\epsilon_0}\frac{dq}{a^2 + z^2}$$

and points directly away from dq, as shown in the figure. We know from symmetry, however, that the net electric field will be along the axis of the ring, so only the component of $d\mathbf{E}$ along this axis will contribute to the net electric field. To extract this component, we need the trig factor

$$\cos\alpha = \frac{z}{\sqrt{a^2 + z^2}}$$

The net electric field at points along the ring's axis will therefore be

$$E_{net} = \frac{1}{4\pi\epsilon_0}\int_{\text{ring}}\frac{dq}{a^2 + z^2}\frac{z}{\sqrt{a^2 + z^2}} \tag{15.10}$$

In general, the expression for dq depends on the geometry of the distribution: for a linear distribution of charge density λ per unit length, dq would be $\lambda\, ds$, where ds is the element of arc; for a surface distribution of charge density σ per unit area, dq would be $\sigma\, dA$, where dA is the area element; and for a volume distribution of charge density ρ, dq would be $\rho\, dV$, where dV is the volume element.[9] Here we have a linear distribution with

$$\lambda = \frac{q}{2\pi a}$$

and, for the element of arc around the perimeter of the ring,

$$ds = a\, d\theta$$

[9]Except that we are dealing with charge rather than mass densities, the relations for dq are the same as those for dm in the center-of-mass calculations of Chapter 6.

Thus [10]

$$dq = \lambda\, ds = \frac{q}{2\pi a}\, a\, d\theta = \frac{q}{2\pi}\, d\theta$$

and eq. (15.10) yields

$$\begin{aligned} E_{net} &= \frac{1}{4\pi\epsilon_0} \int_{\text{ring}} dq\, \frac{z}{(a^2+z^2)^{\frac{3}{2}}} \\ &= \frac{1}{4\pi\epsilon_0} \int_0^{2\pi} \frac{q}{2\pi}\, d\theta\, \frac{z}{(a^2+z^2)^{\frac{3}{2}}} \\ &= \frac{1}{4\pi\epsilon_0}\, \frac{q}{2\pi}\, \frac{z}{(a^2+z^2)^{\frac{3}{2}}} \int_0^{2\pi} d\theta \\ &= \frac{1}{4\pi\epsilon_0}\, \frac{q}{2\pi}\, \frac{z}{(a^2+z^2)^{\frac{3}{2}}}\, (2\pi) \\ &= \frac{1}{4\pi\epsilon_0}\, \frac{qz}{(a^2+z^2)^{\frac{3}{2}}} \end{aligned} \tag{15.11}$$

where we have noted that nothing in the integrand depended on θ and therefore everything could be brought outside the integration.

Now that we have this result, we can see that it makes sense in a couple of limits: as we approach the center of the ring ($z \to 0$), $E_{net} \to 0$, as we would have expected from the symmetry of the distribution. And when we are far away from the ring ($z \gg a$), the a^2 in the denominator is negligible by comparison to the z^2, so that

$$E_{net} \to \frac{1}{4\pi\epsilon_0}\, \frac{qz}{(z^2)^{\frac{3}{2}}} = \frac{1}{4\pi\epsilon_0}\, \frac{q}{z^2}$$

That is, far from the ring the field is the same as that of a point charge, which is also what we would expect: from far away, the ring would look like a point.

Needless to say, you should be able to calculate the electric field of a uniform ring of charge at points along its axis by direct integration.

15.3.2 Disks of Charge

Now consider the case of a uniform disk of charge q and radius a. This is very similar to the case of a ring of charge, except that when we evaluate the electric field on the axis of the disk, at a distance z from its center, the

[10] We could of course equally well have thought in terms of the *angular* distribution of charge: q is uniformly distributed over the 2π angle of the ring, so that the charge per unit angle is $q/2\pi$ and hence $dq = (q/2\pi)\, d\theta$.

charge is not just at the perimeter of the disk; it is spread out from $r = 0$ to $r = a$, so that our squared distance from a bit of charge dq on the disk is now of the more general form $r^2 + z^2$. The electric field of each bit dq of charge is thus of magnitude

$$dE = \frac{1}{4\pi\epsilon_0} \frac{dq}{r^2 + z^2}$$

Again, from symmetry we know that the net electric field will be along the axis of the disk, so that only the component of $d\mathbf{E}$ along the disk's axis will contribute to the net electric field. The trig factor that will extract this component is

$$\frac{z}{\sqrt{r^2 + z^2}}$$

The net electric field at points along the disk's axis will therefore be

$$E_{net} = \frac{1}{4\pi\epsilon_0} \int_{\text{disk}} \frac{dq}{r^2 + z^2} \frac{z}{\sqrt{r^2 + z^2}} \tag{15.12}$$

This time we have a surface distribution of charge per unit area

$$\sigma = \frac{q}{\pi a^2}$$

and we will need to integrate over patches of area

$$dA = r\, dr\, d\theta$$

Thus

$$dq = \frac{q}{\pi a^2} r\, dr\, d\theta$$

and eq. (15.12) yields

$$\begin{aligned} E_{net} &= \frac{1}{4\pi\epsilon_0} \int_{\text{disk}} dq \frac{z}{(r^2 + z^2)^{\frac{3}{2}}} \\ &= \frac{1}{4\pi\epsilon_0} \int_{\text{disk}} \frac{q}{\pi a^2} r\, dr\, d\theta \frac{z}{(r^2 + z^2)^{\frac{3}{2}}} \\ &= \frac{1}{4\pi\epsilon_0} \frac{qz}{\pi a^2} \int_0^a dr \frac{r}{(r^2 + z^2)^{\frac{3}{2}}} \int_0^{2\pi} d\theta \end{aligned}$$

The $d\theta$ integration is trivial. The dr integration is not too bad, either, because of the r in the numerator:

$$\int dr \frac{r}{(r^2 + z^2)^{\frac{3}{2}}} = -\frac{1}{\sqrt{r^2 + z^2}}$$

We therefore have

$$\begin{aligned} E_{net} &= \frac{1}{4\pi\epsilon_0}\,\frac{qz}{\pi a^2}\left[-\frac{1}{\sqrt{r^2+z^2}}\right]_{r=0}^{a}(2\pi) \\ &= \frac{1}{4\pi\epsilon_0}\,\frac{2qz}{a^2}\left(\frac{1}{z}-\frac{1}{\sqrt{a^2+z^2}}\right) \\ &= \frac{1}{4\pi\epsilon_0}\,\frac{2q}{a^2}\left(1-\frac{z}{\sqrt{a^2+z^2}}\right) \end{aligned} \tag{15.13}$$

To see what happens when we are far away from the disk, we note that when $z \gg a$

$$\frac{z}{\sqrt{a^2+z^2}} = \frac{1}{\sqrt{1+(a/z)^2}} \approx \frac{1}{1+\frac{1}{2}(a/z)^2} \approx 1-\frac{1}{2}\Big(\frac{a}{z}\Big)^2$$

so that our result (15.13) reduces to

$$E_{net} \to \frac{1}{4\pi\epsilon_0}\,\frac{2q}{a^2}\left(1-\left(1-\frac{1}{2}\Big(\frac{a}{z}\Big)^2\right)\right) = \frac{1}{4\pi\epsilon_0}\,\frac{q}{z^2}$$

which, as expected, reproduces the field of a point charge.

The limit $z \to 0$ as we go to the center of the disk is a bit trickier: we expect from symmetry that we will get $E_{net} \to 0$, but in fact eq. (15.13) gives

$$E_{net} \to \frac{1}{4\pi\epsilon_0}\,\frac{2q}{a^2}\,(1-0) = \frac{1}{4\pi\epsilon_0}\,\frac{2q}{a^2}$$

If we use $\sigma = q/\pi a^2$ to rewrite this in the form

$$E_{net} \to \frac{1}{4\pi\epsilon_0}\,\frac{2q}{a^2} = \frac{1}{4\pi\epsilon_0}\,2\pi\,\frac{q}{\pi a^2} = \frac{1}{4\pi\epsilon_0}\,2\pi\sigma$$

we see that E_{net} in fact approaches the field of an infinite sheet of charge. This makes sense in that the disk will appear to be of infinite extent as we get very close to it. In fact whenever we cross a surface charge density, that surface, no matter what its shape, will in the limit as we approach it look like an infinite sheet of charge, so that there will always be a finite discontinuity of

$$\frac{1}{4\pi\epsilon_0}\,4\pi\sigma$$

as we go from the infinite-sheet field

$$\frac{1}{4\pi\epsilon_0}\,2\pi\sigma$$

pointing in one direction on one side of the surface to the same field pointing in the opposite direction on the other side of the surface. By symmetry, the

electric field right at the center of our disk must vanish, but that does not mean that the limit as we approach the disk from either side must vanish.

In fact, if we take the limit that the radius of the disk becomes infinite ($a \to \infty$) as the charge density σ remains constant, then

$$\begin{aligned} E_{net} &= \frac{1}{4\pi\epsilon_0}\frac{2q}{a^2}\left(1 - \frac{z}{\sqrt{a^2+z^2}}\right) \\ &= \frac{1}{4\pi\epsilon_0}2\pi\sigma\left(1 - \frac{z}{\sqrt{a^2+z^2}}\right) \\ &\to \frac{1}{4\pi\epsilon_0}2\pi\sigma\,(1-0) \\ &= \frac{1}{4\pi\epsilon_0}2\pi\sigma \end{aligned}$$

which does indeed reproduce the field of an infinite sheet.

You should be able to calculate the electric field of the following uniform distributions of charge by direct integration:

- A disk, at points along the axis of the disk.
- An annulus (that is, a washer), at points along the axis of the annulus.
- An infinite sheet.

15.3.3 Finite Line Segments of Charge

Finally, consider the case of a uniform line segment of charge q and length ℓ. We will evaluate the electric field at points along the perpendicular bisector of the segment, that is, along the dashed blue line of fig. (15.5). The linear charge density on the segment is $\lambda = q/\ell$. If we denote the distance from the segment along this bisector by z and specify points along the segment by a

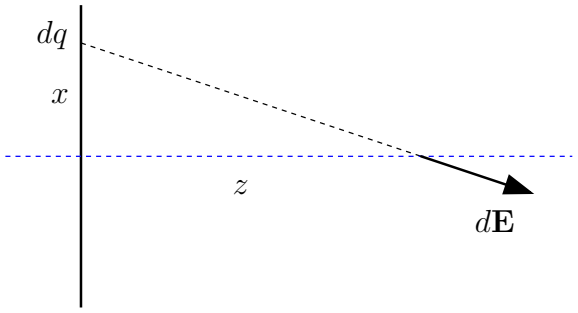


Figure 15.5: Field of a Line Segment of Charge

coordinate x measured from the center of the segment, then the electric field of the charge

$$dq = \lambda\, dx = \frac{q}{\ell}\, dx$$

on the infinitesimal segment dx is of magnitude

$$dE = \frac{1}{4\pi\epsilon_0}\, \frac{dq}{x^2 + z^2} = \frac{1}{4\pi\epsilon_0}\, \frac{q\, dx}{\ell(x^2 + z^2)}$$

By symmetry the net electric field will be along the bisector, and we need the trig factor

$$\frac{z}{\sqrt{x^2 + z^2}}$$

to extract the component of $d\mathbf{E}$ along this axis. The net electric field at points along the bisector will therefore be

$$\begin{aligned} E_{net} &= \frac{1}{4\pi\epsilon_0} \int_{-\frac{1}{2}\ell}^{\frac{1}{2}\ell} \frac{q\, dx}{\ell(x^2 + z^2)}\, \frac{z}{\sqrt{x^2 + z^2}} \\ &= \frac{1}{4\pi\epsilon_0}\, \frac{qz}{\ell} \int_{-\frac{1}{2}\ell}^{\frac{1}{2}\ell} \frac{dx}{(x^2 + z^2)^{\frac{3}{2}}} \\ &= \frac{1}{4\pi\epsilon_0}\, \frac{qz}{\ell}\, 2 \int_{0}^{\frac{1}{2}\ell} \frac{dx}{(x^2 + z^2)^{\frac{3}{2}}} \end{aligned}$$

where in the last step we have noted that, since the integrand is even, the contribution to the integral from $x = -\frac{1}{2}\ell$ to $x = 0$ will be the same as that from $x = 0$ to $x = \frac{1}{2}\ell$.[11] This integration involves a trig substitution that you will not infrequently encounter: the tangent substitution

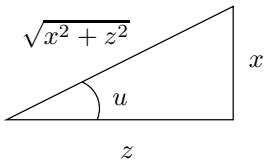
$$x = z \tan u \qquad\qquad dx = z \sec^2 u\, du$$

by which[12]

$$\int \frac{dx}{(x^2 + z^2)^{\frac{3}{2}}} = \int \frac{z \sec^2 u\, du}{(z^2 \tan^2 u + z^2)^{\frac{3}{2}}}$$

[11] There is of course nothing wrong with keeping the range of integration from $-\frac{1}{2}\ell$ to $\frac{1}{2}\ell$, either.

[12] It is of course handy to remember the old triangle trick here: $\tan u = x/z$ is like



and hence $\sin u = x/\sqrt{x^2 + z^2}$, $\cos u = z/\sqrt{x^2 + z^2}$, etc.

$$
\begin{aligned}
&= \frac{1}{z^2} \int du \, \frac{\sec^2 u}{(\tan^2 u + 1)^{\frac{3}{2}}} \\
&= \frac{1}{z^2} \int du \, \frac{\sec^2 u}{(\sec^2 u)^{\frac{3}{2}}} \\
&= \frac{1}{z^2} \int du \, \cos u \\
&= \frac{1}{z^2} \sin u \\
&= \frac{1}{z^2} \frac{x}{\sqrt{x^2 + z^2}}
\end{aligned}
$$

By means of this substitution, we obtain

$$
\begin{aligned}
E_{net} &= \frac{1}{4\pi\epsilon_0} \frac{qz}{\ell} \, 2 \left[\frac{1}{z^2} \frac{x}{\sqrt{x^2 + z^2}} \right]_{x=0}^{\frac{1}{2}\ell} \\
&= \frac{1}{4\pi\epsilon_0} \frac{qz}{\ell} \, 2 \left(\frac{1}{z^2} \frac{\frac{1}{2}\ell}{\sqrt{\frac{1}{4}\ell^2 + z^2}} - 0 \right) \\
&= \frac{1}{4\pi\epsilon_0} \frac{q}{z\sqrt{z^2 + \frac{1}{4}\ell^2}}
\end{aligned}
$$

When we are far away from the line segment ($z \gg \ell$), this reduces to

$$
E_{net} \approx \frac{1}{4\pi\epsilon_0} \frac{q}{z\sqrt{z^2}} = \frac{1}{4\pi\epsilon_0} \frac{q}{z^2}
$$

which, as expected, reproduces the field of a point charge. And if we take the limit that the segment becomes infinite ($\ell \to \infty$) as the charge density λ remains constant, then

$$
\begin{aligned}
E_{net} &= \frac{1}{4\pi\epsilon_0} \frac{q}{z\sqrt{z^2 + \frac{1}{4}\ell^2}} \\
&= \frac{1}{4\pi\epsilon_0} \frac{q}{\ell} \frac{\ell}{z\sqrt{z^2 + \frac{1}{4}\ell^2}} \\
&= \frac{1}{4\pi\epsilon_0} \lambda \frac{\ell}{z\sqrt{z^2 + \frac{1}{4}\ell^2}} \\
&\to \frac{1}{4\pi\epsilon_0} \lambda \frac{\ell}{z\sqrt{\frac{1}{4}\ell^2}} \\
&= \frac{1}{4\pi\epsilon_0} \frac{2\lambda}{z}
\end{aligned}
$$

which reproduces the field of an infinite line of charge.

You should be able to calculate the electric field of the following uniform distributions of charge by direct integration:

- A finite, semi-infinite,[13] or infinite straight line, at points off the line.
- A finite or semi-infinite line segment, at points along the extrapolation of the segment.

15.4 Electric Field Lines

By taking a succession of infinitesimal steps, each in the direction of the net electric field, you can trace out what are known as electric field lines. For a point charge, for example, we have

$$\mathbf{E} = \frac{1}{4\pi\epsilon_0}\frac{q}{r^2}\,\hat{\mathbf{r}}$$

so that our steps would always be in the radial direction: radially outward in the $+\hat{\mathbf{r}}$ direction for a positive charge q and radially inward in the $-\hat{\mathbf{r}}$ direction for a negative charge. The lines that we would trace out in this way are shown in fig. (15.6): the electric field lines of an isolated point charge are radial, emanating outward if the charge is positive and converging inward if the charge is negative.

Fig. (15.7) shows a plot of the electric field lines of a pair of unlike charges of equal magnitude: as you start out close to the positive charge, which we will take, for the sake of definiteness, to be the charge on the left, the field of that charge dominates the net field, so that the field lines are pretty much radially outward from that charge. But as you get far enough from this positive charge that the other charge's contribution to the net field becomes comparable in magnitude, the combination of a field radially outward from the positive charge and radially inward toward the negative charge makes the field lines bend around toward the negative charge. Then as you get close to this negative charge on the right its field dominates the net field, so that the field lines are pretty much radially inward toward that charge. When the charges are equal, all of the field lines that emanate from the positive charge will eventually bend around and converge into the negative charge.[14]

[13] Somehow, it seems that many of you are unfamiliar with semi-infinite line segments. Young people are so out of touch these days. Anyway, a semi-infinite line segment is a line segment that has only one end — the other extends out to infinity. This could, of course, lead to all sorts of interesting philosophical arguments, but this isn't a course in philosophy, and, in spite of any metaphysical angst you may be experiencing, you now know enough

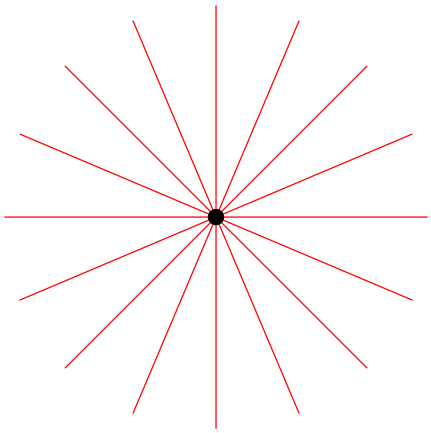


Figure 15.6: Field Lines of a Point Charge

Fig. (15.8) shows the plot of the field lines of a pair of charges equal both in magnitude and sign, which we will take, again for the sake of definiteness, to be two positive charges: the lines start coming pretty much radially outward

to be able to deal with semi-infinite line segments physically and mathematically.

[14]We could worry about the single field line that, starting from the positive charge, goes directly to the left and out to infinity, as well as the corresponding line that comes from out at infinity on the right and goes into the negative charge. But we won't.

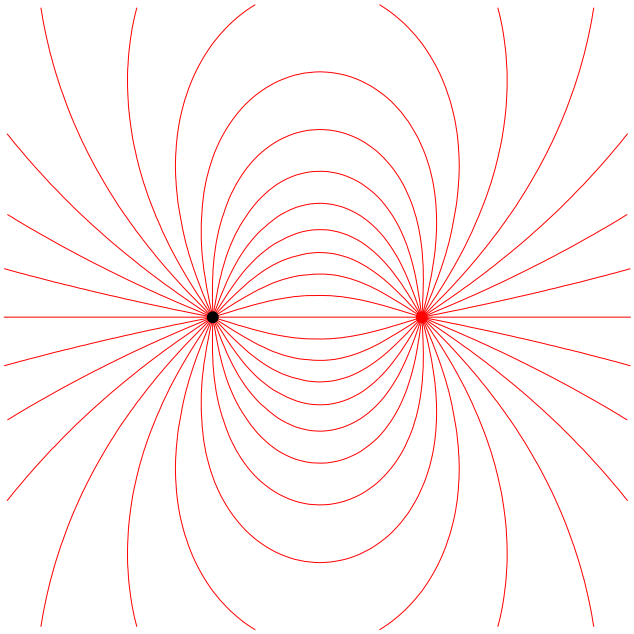


Figure 15.7: Field Lines of Equal and Opposite Charges

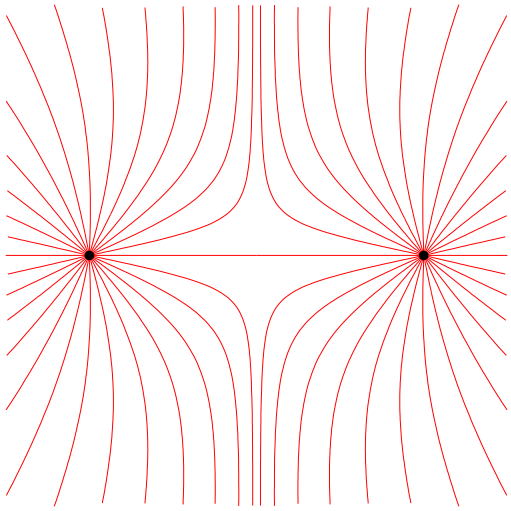


Figure 15.8: Field Lines of Equal and Like Charges

from each charge. But as you get far enough from each charge that the other charge's contribution to the net field becomes comparable in magnitude, the combination of a field radially outward from both charges makes the field lines bend away from both. Very far from the charges — far enough away that the pair of charges would look like a combined point charge at a single point — , the field lines will be coming pretty much radially outward from that point, just like the field lines of a single point charge, as shown in fig. (15.9).

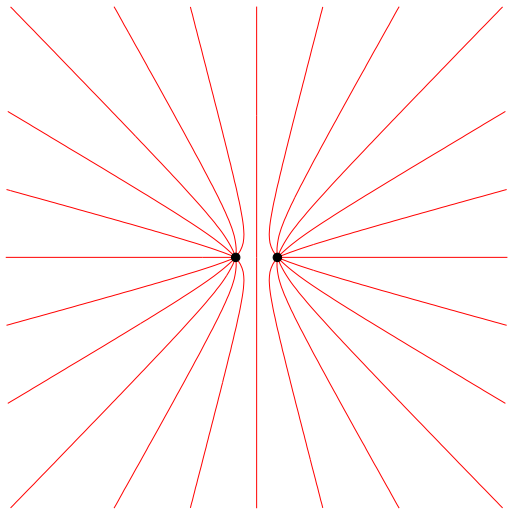


Figure 15.9: Field Lines of Equal and Like Charges

Bear in mind that although we have drawn them in two dimensions, all of these plots of the electric field are really properly three-dimensional, so that the field lines in fig. (15.6), for example, would look like a sea urchin.

What use are these plots? Not much, actually, though for some more complex configurations of charge they can look very pretty. In the back of his massive opus *A Treatise on Electricity and Magnetism*, Maxwell has lots of detailed plots of electric and magnetic fields of various configurations of charge and current, an astounding piece of work when one considers that in those days the calculating and plotting all had to be done by hand. For readers who find them an entertaining diversion, there are more such plots in Appendix F.

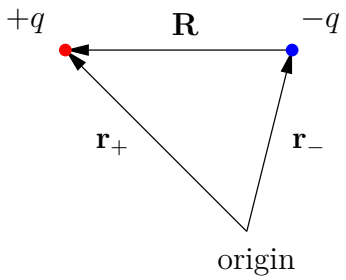


Figure 15.10: An Electric Dipole

15.5 Electric Dipoles

Some molecules, while electrically neutral, are biased, with the orbiting electrons spending more of their time on one side of the molecule than the other. As a result, one side of the molecule effectively carries a net negative charge and the other side an equal and opposite net positive charge. It turns out that to a very good approximation we can account for the electrical properties of such molecules by treating them as a pair of equal and opposite point charges, an arrangement known as an electric dipole. In fig. (15.10), $\mathbf{r}_+$ and $\mathbf{r}_-$ are the position vectors of the positive charge $+q$ and the negative charge $-q$, and $\mathbf{R} = \mathbf{r}_+ - \mathbf{r}_-$ is the vector that goes from $-q$ to $+q$. The dipole moment $\mathbf{p}$ of such an electric dipole is defined to be [15]

$$\mathbf{p} = q\mathbf{R} \tag{15.14}$$

This definition is useful because it turns out that $q\mathbf{R}$ is the quantity involved in both the torque on and potential energy of the dipole: If the dipole is in

[15] Note that, contrary to the that for electric field, the convention for the dipole moment is that $\mathbf{p}$ points *away from* the negative charge and *toward* the positive charge.

an external electric field $\mathbf{E}$, then the net torque on it is

$$\begin{aligned}\boldsymbol{\tau} &= \mathbf{r}_+ \times \mathbf{F}_{on\ +q} + \mathbf{r}_- \times \mathbf{F}_{on\ -q} \\ &= \mathbf{r}_+ \times (q\mathbf{E}) + \mathbf{r}_- \times (-q\mathbf{E}) \\ &= q(\mathbf{r}_+ - \mathbf{r}_-) \times \mathbf{E} \\ &= q\mathbf{R} \times \mathbf{E} \\ &= \mathbf{p} \times \mathbf{E} \end{aligned} \tag{15.15}$$

and its potential energy in this external field is

$$\begin{aligned}U &= -\int_{\mathbf{r}_0}^{\mathbf{r}_+} d\mathbf{r} \cdot \mathbf{F}_{on\ +q} - \int_{\mathbf{r}_0}^{\mathbf{r}_-} d\mathbf{r} \cdot \mathbf{F}_{on\ -q} \\ &= -\int_{\mathbf{r}_0}^{\mathbf{r}_+} d\mathbf{r} \cdot (q\mathbf{E}) - \int_{\mathbf{r}_0}^{\mathbf{r}_-} d\mathbf{r} \cdot (-q\mathbf{E}) \\ &= -\int_{\mathbf{r}_0}^{\mathbf{r}_+} d\mathbf{r} \cdot (q\mathbf{E}) + \int_{\mathbf{r}_0}^{\mathbf{r}_-} d\mathbf{r} \cdot (q\mathbf{E})\end{aligned}$$

which, if we reverse the limits on the second integration and then combine it with the first, becomes

$$\begin{aligned}U &= -\int_{\mathbf{r}_0}^{\mathbf{r}_+} d\mathbf{r} \cdot (q\mathbf{E}) - \int_{\mathbf{r}_-}^{\mathbf{r}_0} d\mathbf{r} \cdot (q\mathbf{E}) \\ &= -\int_{\mathbf{r}_-}^{\mathbf{r}_+} d\mathbf{r} \cdot (q\mathbf{E}) \\ &= -(\mathbf{r}_+ - \mathbf{r}_-) \cdot q\mathbf{E} \\ &= -\mathbf{R} \cdot q\mathbf{E} \\ &= -(q\mathbf{R}) \cdot \mathbf{E} \\ &= -\mathbf{p} \cdot \mathbf{E}\end{aligned} \tag{15.16}$$

Note that the dipole is always presumed to be small in size, so small that even if the external electric field is not constant in magnitude and direction, it will not vary significantly between the two charges of the dipole. We assumed this in our above derivations of the torque on and energy of the dipole when we used the same $\mathbf{E}$ in our expressions for the forces on both of the charges and treated $\mathbf{E}$ as a constant in our line integration between the charges. For this same reason, the net force on the dipole will always vanish:

$$\mathbf{F}_{net} = (+q)\mathbf{E} + (-q)\mathbf{E} = 0 \tag{15.17}$$

From eqq. (15.15) through (15.17) we can discern the behavior of molecular dipoles subjected to an external electric field: Because there is no net force on the dipoles, they do not move in response to the field, but they do

rotate: to the extent that a molecule's dipole moment $\mathbf{p}$ is not aligned with the $\mathbf{E}$ field, there is a torque on it, and from $U = -\mathbf{p} \cdot \mathbf{E}$ we see that the lowest energy state is when the dipole moment $\mathbf{p}$ points in the same direction as $\mathbf{E}$. Thus molecular dipoles tend, modulo the random jostling of thermal motion, to align with an external electric field. This will have consequences for electrical devices known as capacitors in §16.4.5.

We will do a bit more with electrical dipoles, and with a not dissimilar animal known as an electrical quadrupole, in problems #12 and #13.

15.6 Electrostatic Potential & Voltage

The electrostatic potential, also known as the voltage, is the scalar potential φ of eq. (14.19):[16]

$$\mathbf{E} = -\nabla\varphi - \frac{\partial \mathbf{A}}{\partial t}$$

In electrostatics, there is no time dependence, so that the $\partial\mathbf{A}/\partial t$ term vanishes and we have just

$$\mathbf{E} = -\nabla\varphi \tag{15.18}$$

By analogy to

$$\mathbf{F} = -\nabla U \qquad\qquad U = -\int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{F} \cdot d\mathbf{r}$$

we therefore conclude that

$$\varphi = -\int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{E} \cdot d\mathbf{r} \tag{15.19}$$

We will postpone discussion of the physical meaning and uses of the electrostatic potential φ until the following section, where we will see that it is related to potential energy; for the moment we merely note that eq. (15.19) provides us with a technique for determining it: if we happen to know the electric field of a distribution of charge, we can obtain a result for its electrostatic potential by performing the above line integral of $\mathbf{E}$.

For example, suppose we have a uniform solid sphere of charge q and radius a. From eq. (15.3), the sphere's radial electric field is

$$E(r) = \begin{cases} \dfrac{1}{4\pi\epsilon_0}\dfrac{q}{r^2} & (r > a) \\[2ex] \dfrac{1}{4\pi\epsilon_0}\dfrac{qr}{a^3} & (r < a) \end{cases}$$

[16] At the introductory level, most books use V for the electrostatic potential. We will use the big-people symbol φ.

Also note that you now have to be careful about your terminology: when you say simply *potential*, that means the electrostatic potential, a.k.a. voltage; if you mean potential energy then you have to use the full phrase *potential energy*.

Although you can measure the electrostatic potential from wherever you want, it is conventional to measure it from out at infinity.[17] If, for example, we want a result for the potential at the center of the sphere, we would carry out the line integral

$$\varphi = -\int_{r=\infty}^{0} \mathbf{E}\cdot d\mathbf{r}$$

Since we have different expressions for the electric field for $r > a$ and $r < a$, to carry out this integration we need to break it into two separate pieces:

$$\begin{aligned}
\varphi &= -\int_{r=a}^{0} \mathbf{E}\cdot d\mathbf{r} - \int_{r=\infty}^{a} \mathbf{E}\cdot d\mathbf{r} \\
&= -\int_{r=a}^{0} dr\,\frac{1}{4\pi\epsilon_0}\,\frac{qr}{a^3} - \int_{r=\infty}^{a} dr\,\frac{1}{4\pi\epsilon_0}\,\frac{q}{r^2} \\
&= -\frac{1}{4\pi\epsilon_0}\,\frac{q}{a^3}\Big[\tfrac{1}{2}r^2\Big]_a^0 - \frac{1}{4\pi\epsilon_0}\Big[-\frac{q}{r}\Big]_\infty^a \\
&= -\frac{1}{4\pi\epsilon_0}\,\frac{q}{a^3}\Big(-\tfrac{1}{2}a^2\Big) - \frac{1}{4\pi\epsilon_0}\Big(-\frac{q}{a}\Big) \\
&= \frac{1}{4\pi\epsilon_0}\,\frac{3q}{2a}
\end{aligned}$$

More frequently, we will use another method of obtaining the electrostatic potential: Since the field of a point charge q is

$$E = \frac{1}{4\pi\epsilon_0}\,\frac{q}{r^2}$$

the potential at a distance r from q is

$$\begin{aligned}
\varphi &= -\int_{r=\infty}^{r} dr\,\frac{1}{4\pi\epsilon_0}\,\frac{q}{r^2} \\
&= \frac{1}{4\pi\epsilon_0}\,\frac{q}{r}
\end{aligned} \tag{15.20}$$

We may thus obtain the electrostatic potential of a collection of discrete point charges by summing up contributions of the form (15.20) and that of continuous a distribution of charge by integrating over all the point-like infinitesimal charges dq of which it is made up:

$$\varphi = \frac{1}{4\pi\epsilon_0}\int \frac{dq}{r} \tag{15.21}$$

[17] Whenever it is possible to do so, that is; there are configurations of charge for which you must, in order to get a finite result, put the zero of φ at a finite location. But we'll worry about that when we come to it.

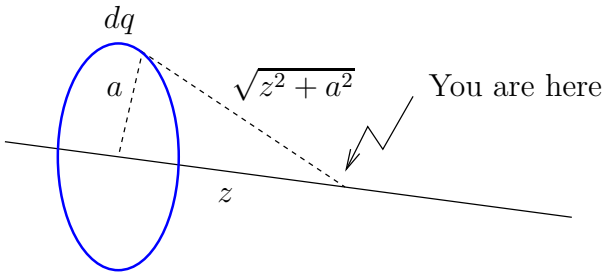


Figure 15.11: Electrostatic Potential of a Ring of Charge

For example, suppose we have a ring of charge q and radius a and that we want to determine the electrostatic potential at points along the axis of the ring (which we will, as usual, parametrize by z). From fig. (15.11), you can see that our distance r from dq is $\sqrt{z^2+a^2}$, so that

$$\varphi = \frac{1}{4\pi\epsilon_0} \int_{\text{ring}} \frac{dq}{\sqrt{z^2+a^2}}$$

Since $\sqrt{z^2+a^2}$ is the same for all of the bits dq of charge around the ring,

$$\begin{aligned} \varphi &= \frac{1}{4\pi\epsilon_0} \frac{1}{\sqrt{z^2+a^2}} \int_{\text{ring}} dq \\ &= \frac{1}{4\pi\epsilon_0} \frac{q}{\sqrt{z^2+a^2}} \end{aligned} \tag{15.22}$$

The electrostatic potential being a scalar rather than a vector, it is often easier first to calculate φ and then use $\mathbf{E} = -\nabla\varphi$ to get a result for the electric field than it is to calculate the electric field directly. Although doing the latter for the ring of charge was not difficult, we can use our result (15.22) to obtain the electric field of the ring at points along its axis: [18]

$$E_z = -(\nabla\varphi)_z = -\frac{\partial}{\partial z}\left(\frac{1}{4\pi\epsilon_0}\frac{q}{\sqrt{z^2+a^2}}\right) = \frac{1}{4\pi\epsilon_0}\frac{qz}{(z^2+a^2)^{\frac{3}{2}}}$$

which is exactly the result we had obtained in §15.3.1.

You should be able to set up and work the electrostatic potential of the following symmetric distributions of charge:

[18] Note that we could not, however, use $\mathbf{E} = -\nabla\varphi$ to obtain the x or y components of $\mathbf{E}$, since by evaluating φ only along the axis of the ring we have already set $x = y = 0$ and thus killed the x and y dependence of φ. It would be like setting $x = 2$ in $f(x)$ and then attempting to take the x derivative of f. (Yeah, you laugh, but we've seen this done before.)

- A solid sphere.
- A spherical shell, of either infinitesimal or finite thickness.
- An infinite solid cylinder.
- An infinite cylindrical shell, of either infinitesimal or finite thickness.
- A finite, semi-infinite, or infinite straight line, at points off the line.
- A finite or semi-infinite line segment, at points along the extrapolation of the segment.
- A thin ring, at points along the axis of the ring.
- A disk or annulus, at points along its axis.

15.7 Equipotential Lines & Surfaces

Plots of the equipotential lines — lines along which the electrostatic potential is constant — are shown in figg. (15.12)-(15.15) for the same configurations of charge for which the electric field was plotted in figg. (15.6)-(15.8). Fig. (15.12) shows the equipotential lines of a point charge: since

$$\varphi = \frac{1}{4\pi\epsilon_0}\,\frac{q}{r}$$

for such a charge, the equipotential lines are lines of constant radius — that is, circles — around the charge. More precisely, since equipotential plots should, like electric field plots, really be three-dimensional, we would have equipotential surfaces consisting of spheres centered on the charge.

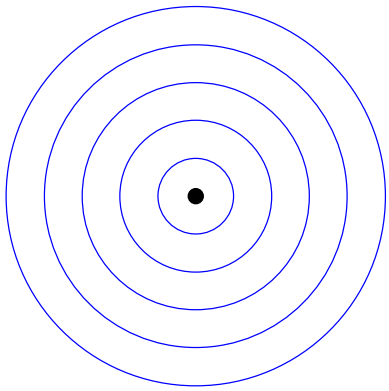


Figure 15.12: Equipotential Lines of a Point Charge

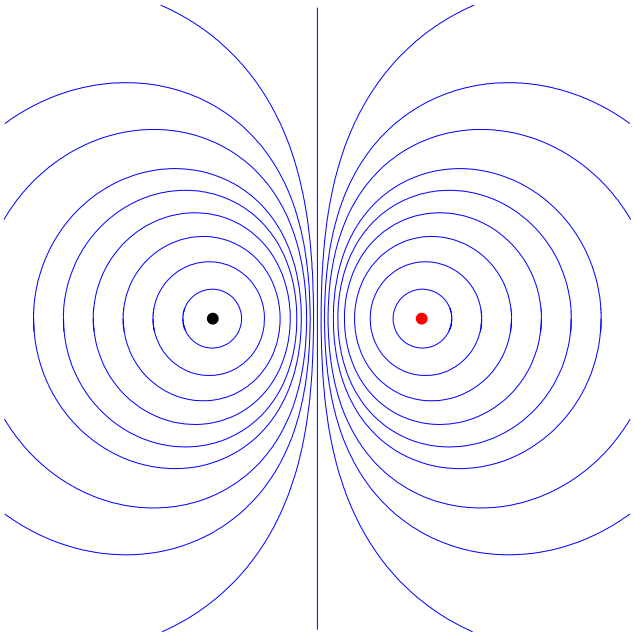


Figure 15.13: Equipotential Lines of Equal & Opposite Charges

Fig. (15.13) shows the equipotential lines of equal and opposite point charges: when we are near one of the charges, its contribution dominates the net electrostatic potential and we therefore pretty much have circles around that charge, but as we get far enough from each charge that the contribution of the other charge becomes significant, these circles become distorted.

Fig. (15.14) shows a similar effect for the case of equal and like point charges: near each charge, we get pretty much circles around that charge, with progressively more distortion as we get into regions where the two charges make comparable contributions to the net electrostatic potential. As you can see from fig. (15.15), as you get farther from the charges the equipotential lines of equal and like point charges merge with each other and eventually become circles around the charges: from far away, the charges will look like a single, combined point charge.

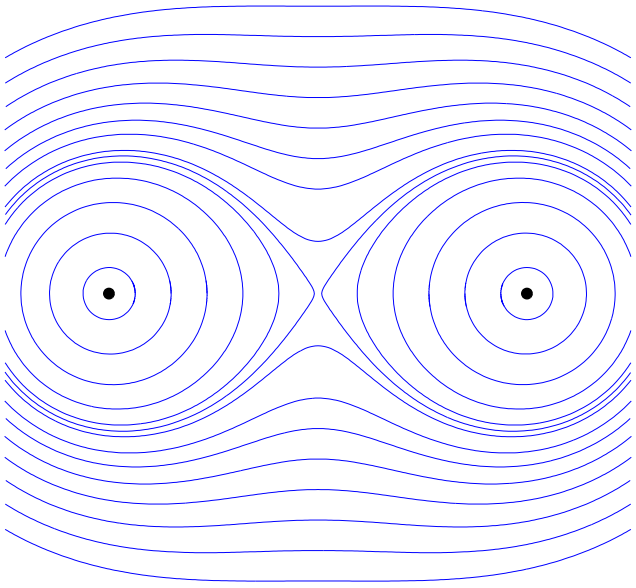


Figure 15.14: Equipotential Lines of Equal & Like Charges

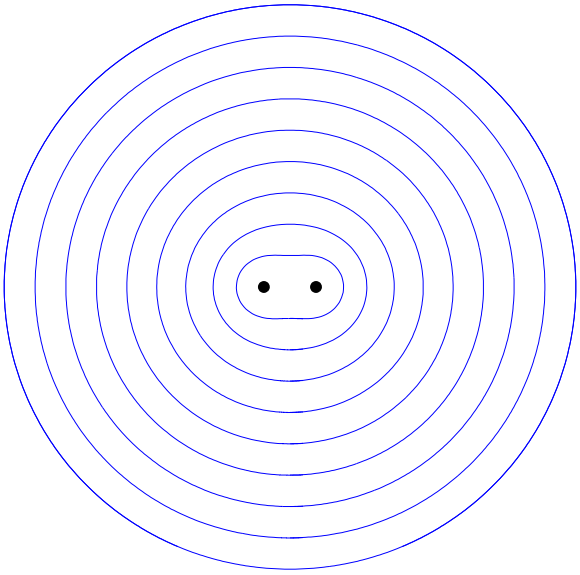


Figure 15.15: Equipotential Lines of Equal & Like Charges

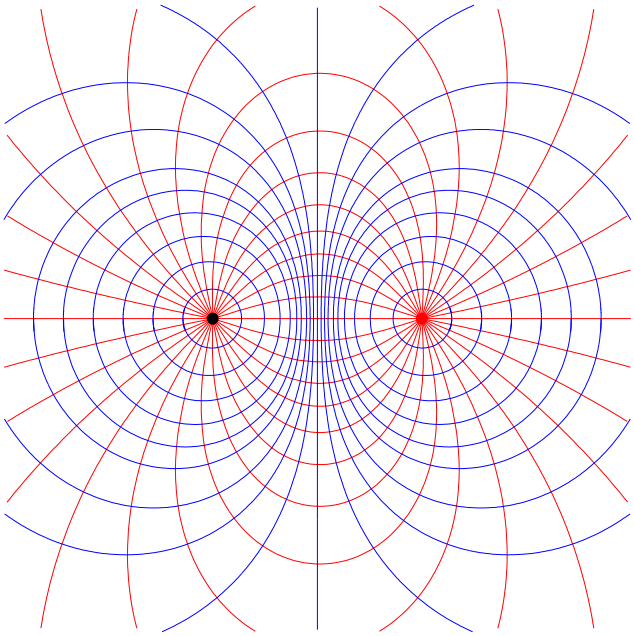


Figure 15.16: Equal and Opposite Charges

As we saw on p.95, the lines and surfaces over which a scalar function U is constant are perpendicular to its gradient ∇U. Since $\mathbf{E} = -\nabla\varphi$, electric field lines are therefore always perpendicular to equipotential lines. For a point charge, for example, the electric field lines are radial and the equipotentials are spheres around the charge. Figg. (15.16)-(15.18) illustrate this effect for more complicated configurations of charge: you can see that the red field lines do indeed always cross the blue equipotential lines at a right angle. Readers with idle time on their hands will find more such plots in Appendix F.

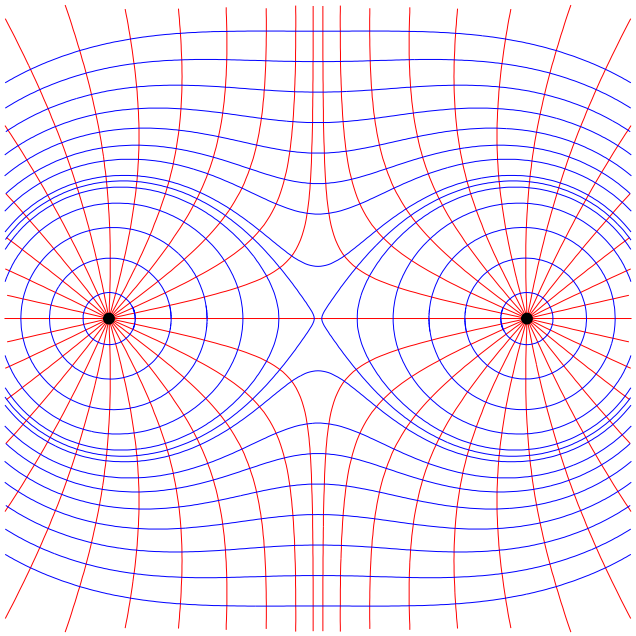


Figure 15.17: Equal and Like Charges

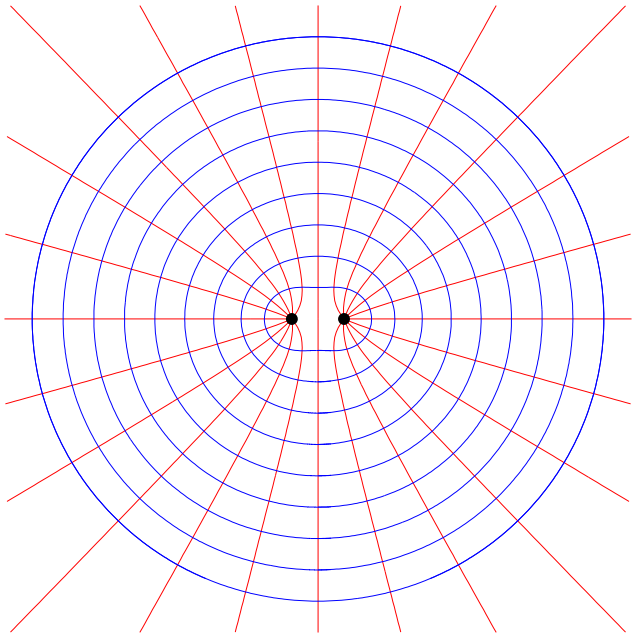


Figure 15.18: Equal and Like Charges

15.8 Electrostatic Potential Energy

Like the potential energy corresponding to any other force, the electrostatic potential energy U corresponding to the electrostatic force $\mathbf{F}$ is given by

$$U = -\int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{F} \cdot d\mathbf{r}$$

Since the force on a point charge q is $\mathbf{F} = q\mathbf{E}$, the electrostatic potential energy of a point charge is

$$U = -\int_{\mathbf{r}_0}^{\mathbf{r}} q\mathbf{E} \cdot d\mathbf{r} = q\Big(-\int_{\mathbf{r}_0}^{\mathbf{r}} \mathbf{E} \cdot d\mathbf{r}\Big)$$

which, if we use the definition (15.19) of electrostatic potential, reduces to

$$U = q\varphi \tag{15.23}$$

That is, the electrostatic potential energy of a point charge is just the product of the charge with the electrostatic potential at its location. The electrostatic potential φ may therefore be thought of as the energy per unit charge.

When we are dealing with a pair of point charges q_1 and q_2, the electrostatic potential of q_2 is, by eq. (15.20),

$$\varphi_2 = \frac{1}{4\pi\epsilon_0}\,\frac{q_2}{r}$$

The corresponding electrostatic potential energy of the charge q_1 — or what we would refer to simply as the electrostatic potential energy of the pair of charges — is thus

$$\begin{aligned} U &= q_1\varphi_2 \\ &= q_1\Big(\frac{1}{4\pi\epsilon_0}\,\frac{q_2}{r}\Big) \\ &= \frac{1}{4\pi\epsilon_0}\,\frac{q_1 q_2}{r} \end{aligned} \tag{15.24}$$

For configurations of several point charges, the total potential energy will be the sum of the potential energies of each of the *distinct* pairs. So if we have three charges, the total electrostatic potential energy will be

$$U = \frac{1}{4\pi\epsilon_0}\left(\frac{q_1 q_2}{r_{12}} + \frac{q_2 q_3}{r_{23}} + \frac{q_1 q_3}{r_{13}}\right)$$

where r_{ij} is the distance between q_i and q_j.

The potential energy of a continuous distribution of charge can be calculated by integrating contributions analogous to those of eq. (15.24) for the

point-like infinitesimal charges dq that make up the distribution. In the general case, this formally requires performing a double integral like

$$U = \tfrac{1}{2} \iint \frac{dq\, dq'}{|\mathbf{r} - \mathbf{r}'|} = \tfrac{1}{2} \int dV \int dV' \, \frac{\rho(\mathbf{r})\, \rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}$$

where we have used $dq = \rho\, dV$, and where the factor of $\frac{1}{2}$ is to correct for double counting: the integrations by dq and dq' are both over the entire distribution, and as we perform these integrations we do not want to redundantly count both the contribution that pairs up the dq at location B with the dq' at location A and the contribution that pairs up the dq at location A with the dq' at location B.

While performing such double integrals, which in practice are usually prohibitively difficult to carry out, is beyond the scope of this course, there are ways of dealing with simple, highly symmetric distributions of charge. Consider, for example, a uniform solid sphere of charge q and radius a. To obtain a result for the electrostatic potential energy of the sphere we imagine constructing it out of infinitesimally thick spherical shells, each brought in from out at infinity. If a sphere of radius r has already been so constructed, then the electrostatic potential energy at its surface is

$$\varphi = \frac{1}{4\pi\epsilon_0} \frac{q_{so\ far}}{r}$$

where the charge $q_{so\ far}$ that has so far been built up, being proportional to the volume that has so far been built up, is

$$q_{so\ far} = \frac{\frac{4}{3}\pi r^3}{\frac{4}{3}\pi a^3}\, q = \left(\frac{r}{a}\right)^3 q$$

Thus we have

$$\varphi = \frac{1}{4\pi\epsilon_0} \left(\frac{r}{a}\right)^3 q\, \frac{1}{r} = \frac{1}{4\pi\epsilon_0} \frac{qr^2}{a^3}$$

The next infinitesimal shell to be brought in is of area $4\pi r^2$ and thickness dr, so that it has a volume of

$$dV = 4\pi r^2\, dr$$

and therefore carries a charge

$$dq = \rho\, dV = \frac{q}{\frac{4}{3}\pi a^3}\, 4\pi r^2\, dr = \frac{3q}{a^3}\, r^2\, dr$$

The change in the sphere's electrostatic potential energy is thus

$$dU = dq\, \varphi = \left(\frac{3q}{a^3}\, r^2\, dr\right) \left(\frac{1}{4\pi\epsilon_0} \frac{qr^2}{a^3}\right) = \frac{1}{4\pi\epsilon_0} \frac{3q^2}{a^6}\, r^4\, dr$$

If we integrate this starting from nothing ($r = 0$) out to the full sphere of charge ($r = a$), we arrive at

$$U = \int dU = \int_0^a \frac{1}{4\pi\epsilon_0}\frac{3q^2}{a^6}\,r^4\,dr = \frac{1}{4\pi\epsilon_0}\frac{3q^2}{a^6}\frac{a^5}{5} = \frac{1}{4\pi\epsilon_0}\frac{3q^2}{5a} \tag{15.25}$$

Just in case you were curious. We will not, however, being doing anything further with the potential energy of continuous distributions of charge.[19]

15.9 Conductors

A conductor is a substance in which some of the electrons, called conduction electrons, are not bound to specific atoms or molecules and are therefore more or less free to roam about the substance. It is because of the relative freedom of motion of these conduction electrons that conductors so readily carry electric currents.

Conductors have the following properties:

- The electric field vanishes everywhere within the body of a conductor.[20]

  To see why this is so, suppose that there were some nonzero electric field $\mathbf{E}$ within a conductor. By $\mathbf{F} = q\mathbf{E}$, there would then be an electric force on the conduction electrons, which would move in response to that force. The conduction electrons would keep moving until they had redistributed themselves so that there was no longer any force on them and therefore no electric field. In other words, the conduction

[19] It is perhaps worth noting that, the electrostatic and Newtonian gravitational forces both being inverse-square laws, we can obtain a result for the gravitational potential energy of a uniform solid sphere of mass directly from (15.25): comparing

$$F = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r^2} \qquad\qquad F = \frac{Gm_1 m_2}{r^2}$$

we see that we should set

$$\frac{1}{4\pi\epsilon_0} \to G \qquad\qquad q^2 \to m^2$$

We also need to remember that, whereas the electrostatic force is a repulsion, the gravitational force is an attraction, which means a sign difference when we evaluate $U = -\int \mathbf{F}\cdot d\mathbf{r}$. For the sphere of mass, we therefore have

$$U = -\frac{3Gm^2}{5a}$$

[20] By *within the body of the conductor* we mean those points within the substance of the conductor itself, excluding any hollow regions in its interior. The electric field need not vanish in such cavities.

electrons automatically redistribute themselves in such a way that their own electric field cancels out (at least within the body of the conductor) any electric field externally applied to the conductor.[21]

- As a direct consequence of the preceding property, there is no *net* charge anywhere within the body of a conductor; any net charge on a conductor will be distributed over its surface (or surfaces).[22]

  By Gauss's law

  $$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\,\rho \tag{14.1a}$$

  the vanishing of the electric field $\mathbf{E}$ everywhere within the body of the conductor means that the net charge density ρ also vanishes. There being no place else it can go, any net charge that you put on a conductor will therefore distribute itself in some manner over its surface or surfaces.

  Note that there is, of course, still charge within the body of the conductor; it is just that within the conductor's body, the negative charge of the electrons and the positive charge of the nuclei are balanced.

- Another consequence of the vanishing of the electric field within a conductor is that a conductor is an equipotential — that is, all points within the body of the conductor are at the same electrostatic potential (voltage).

  Recall eq. (15.19),

  $$\varphi = -\int_{\mathbf{r}_0}^{\mathbf{r}} d\mathbf{r} \cdot \mathbf{E}$$

  If $\mathbf{r}_1$ and $\mathbf{r}_2$ are any two points within the conductor, and we integrate over a path also within the conductor, the potential difference between these two points is therefore

  $$\Delta\varphi = -\int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{E} \cdot d\mathbf{r}$$

  Since $\mathbf{E} = 0$ everywhere within the conductor, $\Delta\varphi = 0$: there is no potential difference between any two points in the conductor.

[21] In principle it is possible to apply a strong enough external electric field that the conduction electrons would be unable to completely cancel the external field within the body of the conductor, but this really isn't an issue in practice.

[22] Again, by *within the body of the conductor*, we mean those points within the substance of the conductor itself, excluding any hollow regions in its interior. In the case of a hollow sphere, for example, there cannot be any net charge within the spherical conducting shell itself, but there can be net charge outside the sphere, within the hollow region enclosed by the sphere, and on both the inner and outer surfaces of the sphere.

- The electric field immediately outside the surface of a conductor is perpendicular to the surface and of magnitude $E = \sigma/\epsilon_0$.

  Since the entire conductor, and therefore in particular its surface, is an equipotential, and since, as we established on p.805, the electric field is always perpendicular to equipotentials, the electric field at the surface of a conductor must be perpendicular to the surface.

  To determine the magnitude of the field, we apply Gauss's law

  $$\oint_S dA\,\hat{\mathbf{n}} \cdot \mathbf{E} = \frac{1}{\epsilon_0}\, q_{enclosed} \tag{14.2a}$$

  in much the same way that we did for infinite planar sheets of charge in §15.1.3, taking as our Gaussian surface an infinitesimal pillbox that intersects the surface of the conductor. Since the electric field is perpendicular to the surface of the conductor, the contribution to

  $$\oint dA\,\hat{\mathbf{n}} \cdot \mathbf{E}$$

  from the side of this pillbox vanishes. There is also no contribution from the end-cap that lies inside the conductor, because $E = 0$ within the conductor. The only nonzero contribution comes from the outer end-cap of the pillbox. Since the pillbox is infinitesimal, the surface of the conductor will, no matter how funky in shape or charge distribution on a finite scale, seem like an infinite planar sheet of uniform charge density from the perspective of the pillbox, so that the electric field will be constant in magnitude over the outer end-cap as well as perpendicular to it. The left-hand side of Gauss's law thus reduces to $E\,dA$ for the pillbox, where dA is the area of an end-cap. And on the right-hand side of Gauss's law, the charge enclosed is just the charge $\sigma\,dA$ intersected by the pillbox. Thus we arrive at

  $$E\,dA = \frac{1}{\epsilon_0}\,\sigma\,dA$$

  which yields $E = \sigma/\epsilon_0$, as advertised.

For the construction of electrical circuits, the equipotentiality of conductor's is one of their most important properties. When, for example, you want to light a light-bulb, it wouldn't be very practical to have to bring the terminals of electrical generator into direct contact with the bulb, so you need some way of conveying the voltages at the generator's terminals to the bulb. A two-strand conducting wire will do just what you want: when the strands are connected to the terminals of the generator, the voltages of the strands are set to the voltages of the generator's terminals and are then, because

all of a conducting strand must be at the same voltage, effectively conveyed down the length of each strand to its far end, where it can be connected to the bulb. In fact, the power plant that is the ultimate source of the electricity used by your household appliances is — if you're lucky — miles away from your home; the voltage generated at the power plant is conveyed to your household outlets and hence to your appliances by equally many miles of conducting wire.

There is actually a continuous spectrum of conductivity; it is a property that all substances have in some degree, even if that degree is minuscule. What we usually mean by the term conductor is a substance that conducts *well*. It turns out that in metals typically one or two of the outermost electrons in each metal atom are not bound to just that atom; they can roam more or less freely throughout the metal lattice from one atom to another. This is why metals are usually good conductors. There is also a special quantum-mechanical effect manifested by some substances at low temperatures. These substances, known as superconductors, may conduct electricity only very poorly at higher temperatures, but once they are cooled to a certain critical temperature (usually around the temperature of liquid nitrogen or even lower), they suddenly become *perfect* conductors: the conduction electrons in superconductors are totally free to move without any resistance at all.

15.10 The Method of Images

The properties of conductors discussed in the preceding section can sometimes be used to determine electric fields and charge distributions. Consider, for example, a semi-infinite conducting slab with a point charge q a perpendicular distance z from the slab, as shown, albeit with decidedly finite dimensions, in fig. (15.19). Physically, we expect that the presence of the

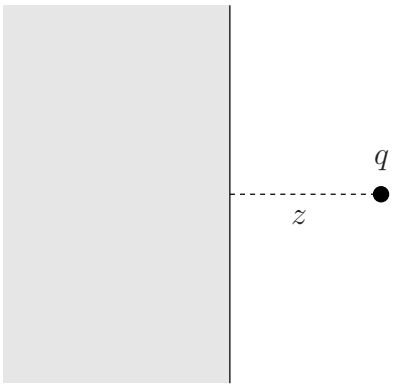


Figure 15.19: The Old Point Charge & Semi-Infinite Slab

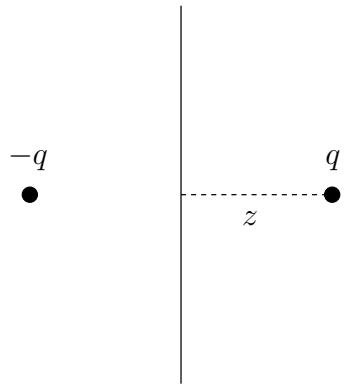


Figure 15.20: The Mirror-Image Point Charge

point charge q will, by attraction, draw opposite charge to the planar surface of the slab, with the result that there will be some net surface charge density σ distributed over it. We also know that since the conducting slab is an equipotential, at points on its surface the net electric field must be perpendicular to the surface — a feature which should bring to mind the electric field of equal and opposite point charges shown in fig. (15.7) on p.795. This leads us to suspect that the surface charge distribution σ on the surface of the slab will mimic the effect we would get from a mirror-image point charge $-q$ inside the slab, as shown in fig. (15.20). If we measure our radii r from the point where the line between q and $-q$ crosses the slab's surface, the net electric field at points on the surface of the slab is then (see fig. (15.21))

$$E_{net} = 2E_q \sin\theta$$

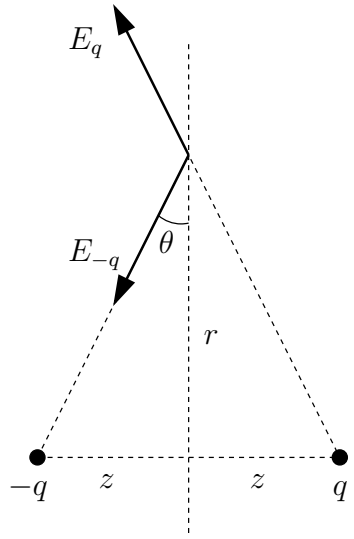


Figure 15.21: The Geometry of the Net Electric Field

$$
\begin{aligned}
&= 2\Big(\frac{1}{4\pi\epsilon_0}\,\frac{q}{r^2+z^2}\Big)\sin\theta \\
&= 2\Big(\frac{1}{4\pi\epsilon_0}\,\frac{q}{r^2+z^2}\Big)\,\frac{z}{\sqrt{r^2+z^2}} \\
&= \frac{1}{4\pi\epsilon_0}\,\frac{2qz}{(r^2+z^2)^{\frac{3}{2}}}
\end{aligned}
\tag{15.26}
$$

To get the net electric field at other locations outside the slab, we would carry out a similar vector sum of the fields of the two point charges. (Inside the slab $E = 0$, of course.) And since the field at the surface of a conductor is σ/ϵ_0, for the charge density σ on the slab's surface eq. (15.26) gives us [23]

$$\sigma = -\frac{1}{4\pi}\,\frac{2qz}{(r^2+z^2)^{\frac{3}{2}}}$$

[23] We have noted that this surface charge will of course be opposite in sign to the charge q.

15.11 Problems

1. In a hydrogen atom, the electron may be crudely regarded as orbiting circularly around the proton at a radius of 5.3×10^{-11} m.

   (a) Compare the magnitudes of the gravitational and electric forces that the electron and proton exert on each other.

   (On the assumption that you're too **** lazy to look them up, the mass of the electron is approximately 9.11×10^{-31} kg, the mass of the proton 1.67×10^{-27} kg, and the magnitude of the charge on the electron and proton 1.60×10^{-19} C.)

   (b) How do you reconcile this with your everyday experience, in which the gravitational force seems so much stronger to the electric force?

2. You put on rubber gloves and charge up two identical cute, furry kittens with static electricity by rubbing them vigorously against a glass surface. You smash and grind the kittens together to be sure that they carry equal charges, then mount them on the ends of wooden poles. You find that the kittens repel each other with a force F_0 when they are held a distance ℓ apart. What can you conclude about the magnitude and sign of the charge on each kitten?

3. How do we know that electrons are negatively charged and protons positively charged? Why not the other way around?

4. What velocity does an electron (charge $-e$, mass m_e) acquire as a result of being accelerated, starting essentially from rest, through a potential difference (that is, voltage) V_0?

   (It is customary to denote by e the absolute value of the fundamental charge — the charge carried by the electron and proton. Thus the charge on the proton is $+e = 1.60217653 \times 10^{-19}$ C and that on the electron $-e$.)[24]

[24] Not everyone adheres to this convention, however; sometimes e is used to denote the electron charge, including the sign, so that $e < 0$. This of course leads to considerable confusion. But at least it hasn't started any wars.

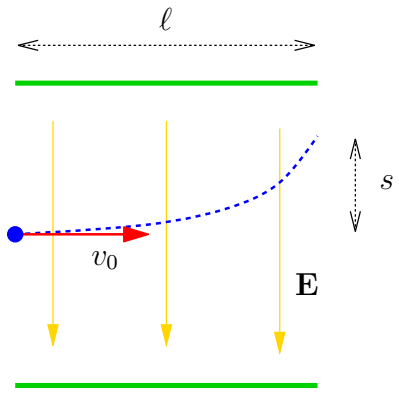


Figure 15.22: Problem 5

5. As shown in fig. (15.22) with a gratuitous and perhaps even offensive use of color, an electron is projected with an initial speed v_0 into the space between two charged metal plates (the edges of which, in a view from the side, are indicated by the green lines in the figure). The charge on the plates is such that everywhere in between them there is a constant downward electric field **E** (the gold arrows). As the blue electron travels the length ℓ of the plates, it follows the deflected trajectory indicated by the dotted blue line, undergoing a displacement s in the perpendicular direction.[25] Determine the value of the electric field between the plates.

[25]Betcha didn't know electric fields were gold and electrons were blue, did you? Or maybe we're just being a little synæsthetic.

Anyway, this is the mechanism by which CRT (cathode ray tube) monitors and TVs work: A beam of electrons that has been accelerated through a potential difference (as in problem #4) enters a region enclosed by two pairs of plates, one pair horizontal and the other pair vertical. The horizontal and vertical components of the electric field in this region are controlled by varying the voltages across the two pairs of plates. These voltages, and hence the deflection of the beam, can be varied quickly and precisely enough to make the beam scan the entire screen in a series of slightly jagged horizontal lines, wagging up and down to illuminate each of the phosphorescent pixels on the screen to the desired degree.

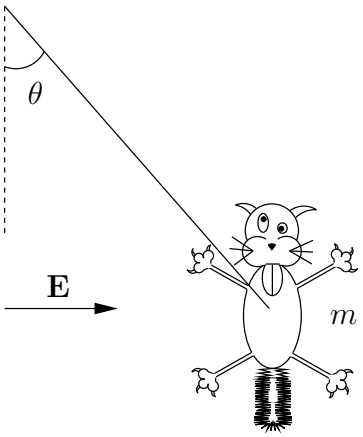


Figure 15.23: Problem 6

6. A standard test cat of mass m is charged up with a Van de Graf generator [26] and, when suspended in a uniform electric field of magnitude E directed toward the right, makes an angle θ with the vertical, as shown in fig. (15.23).

   Determine the charge q on the cat and, if possible, its sign.

7. (Kitty déjà vu.) You affix a cute, furry little kitten (mass m, effectively a sphere of radius a) to the end of a wooden pole of length ℓ and charge it up with a Van de Graf generator[26] to voltage V_0.

   (a) Determine the charge on the kitten.

   (b) If you charge a second, identical kitten to the same voltage and then hold the two kittens a distance $R \gg a$ apart, what is the electrostatic potential energy of the kittens?

   (c) How much work did *you* have to do to assemble this arrangement (that is, to bring the kittens from essentially infinitely far apart to their final state, in which they are separated by distance R)?

   (d) How would your answers to the preceding parts change if the second kitten were instead charged to voltage $-V_0$?

   (e) If the separation R of the kittens were comparable to a, how would the kittens' electrostatic potential energy differ from your result for #7b?

[26] In case you're unfamiliar with these beasts: they consist essentially of a metal dome in contact, by means of a metal brush, with a little conveyor belt that, by rubbing against a suitable material at its other end, conveys static electric charge onto the dome. And the point of this Rube-Goldberg-like arrangement is to build up substantial charges and voltages on the dome — charges and voltages that can then be used to zap the unsuspecting. Remind me to get one out so we can "play" with it. Some years we've run hidden wires from the Van de Graf to a chair in the back of the room, with spectacular effect. *Bwahahahaha!*

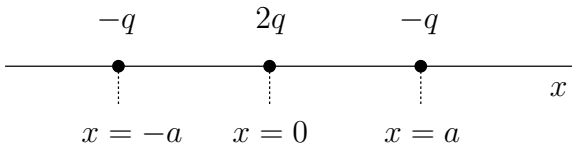


Figure 15.24: Problem 8

8. (Warning: This problem is excruciatingly boring. But it hurts us more than it hurts you.)

   Some stupid positive point charge $2q$ is placed at the origin of an x axis. Two other stupid point charges, each $-q$, are placed at $x = -a$ and $x = +a$, as shown in fig. (15.24).[27]

   (a) What are the magnitude and direction of the net electric force on the charge at the origin?

   (b) What are the magnitude and direction of the net electric force on the charge at $x = a$?

   (c) What are the magnitude and direction of the net electric force on the charge at $x = -a$?

   (d) You should find that the net electric forces on each of the three charges together add up to zero. What is the physical (as opposed to the algebraic) reason for this?

   (e) What are the magnitude and direction of the net electric field at $x = a$ due to the charges at $x = -a$ and $x = 0$?

   (f) What are the magnitude and direction of the net electric field at $x = -2a$?

   (g) What is the total potential energy of the arrangement of the three charges?

   (h) What is the electrostatic potential at $x = -2a$?

   (i) How much work would you have to do to move a point charge Q from infinitely far away to $x = -2a$?

   (j) ... Okay, maybe we'll stop here.

[27] This arrangement is the simplest realization of what is known as an electric quadrupole. Not that that will significantly improve your quality of life. Or make the problem any more exciting.

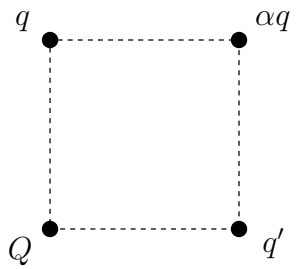


Figure 15.25: Problem 9

9. Four point charges, q, q', Q, and αq are arranged rather lamely at the corners of a square of side a, as shown in fig. (15.25). The parameters q, q', and Q are positive, but α may be positive, negative, or zero.

   (a) If $\alpha = 2$ and $q' = q$, determine

      i. The magnitude and direction of the net electric force on Q.

      ii. The net electric field at the location of Q.

   (b) If $q' = q$, are there values of α for which the net force on Q will vanish? If not, why not? If so, determine those values.

   (c) If $q' = -q$, are there values of α for which the net force on Q will vanish? If not, why not? If so, determine those values.

Figure 15.26: Problem 10

10. Fig. (15.26) shows two point charges on an x axis, a positive charge q at the origin and a charge $-16q$ at $x = \ell$.

    (a) Are there locations x at which a third charge Q can be placed such that the *system* of the three charges is in equilibrium? If not, why not? If so, determine all the possible values of x and Q that will produce equilibrium.

    (b) Are your equilibrium solutions for the preceding part, if any, stable or unstable?

    (c) Okay, let's end this charade: there is an equilibrium solution, and for such a solution the forces must of course be balanced on each of the three charges, for a total of three algebraic conditions. There are, however, only two unknowns. How do you know that this system of three equations and two unknowns must be consistent? Or, in other words, how do you know that if the forces are balanced on two of the three charges, they must also be balanced on the third charge?

    (d) Are there any locations off the x axis where a charge Q would result in the system being in equilibrium? If not, why not? If so, determine all the locations and values of Q that will produce equilibrium.

Figure 15.27: Problem 11

11. Fig. (15.27) shows two point charges on an x axis, a positive charge q at the origin and a charge $-4q$ at $x = \ell$.

    (a) Determine all finite locations along the x axis where the net electrostatic potential vanishes, assuming that the potential is taken to vanish out at infinity.

    (b) Determine all finite locations along the x axis where the net electric field vanishes.

    (c) Determine all finite locations along the x axis where the total electrostatic potential energy vanishes.

    (d) Are there places off the x axis where the net electrostatic potential vanishes? If not, why not? If so, determine where.

Figure 15.28: Problem 12

12. Fig. (15.28) shows a simple electric dipole consisting of equal and opposite point charges, q and $-q$, where q is positive. We will take these charges to be located at $(0, a)$ and $(0, -a)$, respectively.

    (a) Determine the net electric field (both magnitude and direction) as a function of x at locations along the x axis.

    (b) Show that for large x the magnitude of the net electric field is

    $$E \approx \frac{1}{4\pi\epsilon_0}\,\frac{2qa}{x^3} = \frac{1}{4\pi\epsilon_0}\,\frac{p}{x^3}$$

    where, according to eq. (15.14) on p.797, $p = 2qa$ is the dipole moment.

    (c) Determine the electrostatic potential as a function of x at locations along the x axis.

    (d) Why can't you obtain your result for #12a from your result #12c by

    $$\mathbf{E} = -\nabla\varphi \quad ?$$

    (e) Can you obtain your result for #12c from your result for #12a

    $$\varphi = -\int \mathbf{E}\cdot d\mathbf{r} \quad ?$$

(Continued on next page . . .)

Figure 15.28: Problem 12

(f) Determine the net electric field (both magnitude and direction) as a function of y at locations along the y axis outside the dipole (that is, for $y > a$ or $y < -a$).

(g) Show that for large y the magnitude of the net electric field is

$$E \approx \frac{1}{4\pi\epsilon_0}\frac{4qa}{y^3} = \frac{1}{4\pi\epsilon_0}\frac{2p}{y^3}$$

(h) Determine the electrostatic potential as a function of y at locations along the y axis outside the dipole (that is, for $y > a$ or $y < -a$).

(i) Show that you can obtain your result for #12f from your result #12h by

$$\mathbf{E} = -\nabla\varphi$$

(j) Show that you can obtain your result for #12h from your result for #12f

$$\varphi = -\int \mathbf{E}\cdot d\mathbf{r}$$

Now that you've survived all that, note that whereas the electrostatic potential and field of a point charge (monopole) fall off as $1/r$ and $1/r^2$, respectively, those of the electric dipole fall off as $1/r^2$ and $1/r^3$: because there is no net charge on the electric dipole, the equal and opposite monopole contributions cancel each other out, and what remains turns out to fall off one power of r more rapidly. (We have shown this only for points on the x and y axes, but it holds true at all points (x, y, z) outside the dipole.)

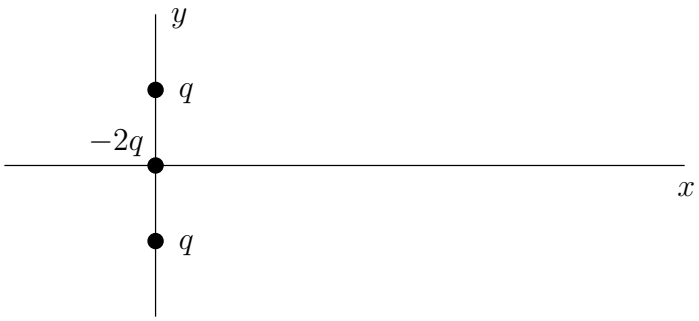


Figure 15.29: Problem 13

13. Fig. (15.29) shows a simple electric quadrupole consisting of two positive point charges q located at $(0, a)$ and $(0, -a)$ and a point charge $-2q$ located at the origin.

    (a) Determine the net electric field (both magnitude and direction) as a function of x at locations along the x axis.

    (b) Show that for large x the magnitude of the net electric field is

    $$E \approx \frac{1}{4\pi\epsilon_0}\frac{3qa^2}{x^4}$$

    (c) Determine the electrostatic potential as a function of x at locations along the x axis.

    (d) Show that you can obtain your result for #13a from your result #13c by

    $$\mathbf{E} = -\nabla\varphi$$

    (e) Show that you can obtain your result for #13c from your result for #13a

    $$\varphi = -\int \mathbf{E}\cdot d\mathbf{r}$$

    (f) Determine the net electric field (both magnitude and direction) as a function of y at locations along the y axis outside the quadrupole (that is, for $y > a$ or $y < -a$).

    (g) Show that for large y the magnitude of the net electric field is

    $$E \approx \frac{1}{4\pi\epsilon_0}\frac{6qa^2}{y^4}$$

(Continued on next page . . .)

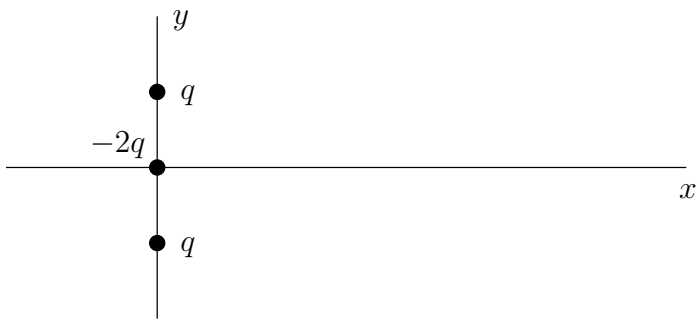


Figure 15.29: Problem 13

(h) Determine the electrostatic potential as a function of y at locations along the y axis outside the dipole (that is, for $y > a$ or $y < -a$).

(i) Show that you can obtain your result for #13f from your result #13h by

$$\mathbf{E} = -\nabla\varphi$$

(j) Show that you can obtain your result for #13h from your result for #13f

$$\varphi = -\int \mathbf{E} \cdot d\mathbf{r}$$

Note that just as an electric dipole consists of equal and opposite charges (monopoles), an electric quadrupole can be regarded as equal and opposite dipoles: the arrangement shown in fig. (15.29) can be constructed by superposing two dipoles of the kind shown in fig. (15.28) facing opposite directions. And just as the cancellation of the monopole contributions resulted in the dipole potential and field falling off as $1/r^2$ and $1/r^3$, respectively, the cancellation of both the monopole and dipole contributions results in the quadrupole potential and field falling off as $1/r^3$ and $1/r^4$. (Again, we have shown this only for points on the x and y axes, but it holds true at all points (x, y, z) outside the quadrupole.)

In fact, we can always construct the next higher multipole by superposing two multipoles facing opposite ways, with the result that the cancellation of the contributions from all of the lower multipoles leaves us with a potential and field that fall off one power of r faster. By this scheme, a monopole involves a single charge, a dipole two charges, a quadrupole $2 \times 2 = 4$ charges, and octopole $2 \times 4 = 8$ charges, and so on. Hence the names *monopole*, *dipole*, *quadrupole*, and *octopole*: since the nth multipole will involve 2^{n-1} charges, it is called a 2^{n-1}-pole. And, as we have seen for the dipole and quadrupole, the potential and field of a 2^{n-1}-pole fall off as $1/r^n$ and $1/r^{n+1}$.

14. A positive point charge q is at the center of a tetrahedron, each edge of which has length ℓ. Determine the flux through each face of the tetrahedron. See the footnote if you need a hint.[28]

15. Explain why you can't use Gauss's law to determine the electric field of a cube, even if the charge within the cube is evenly distributed.

16. Earnshaw's theorem states that it is impossible to construct a completely stable equilibrium point with electrostatic forces alone. Prove this by applying Gauss's law to an infinitesimal charge-free region that contains the supposed equilibrium point. See the footnote if you need a hint.[29]

17. (a) Suppose a positive point charge q is brought near (but not in contact with) a solid, electrically neutral conducting sphere. Describe qualitatively how charge will distribute itself within and on the surface of the sphere. (Note that although the stipulation *electrically neutral* means zero *net* charge on the sphere, by virtue of the sphere's being a conductor at least some of the positive and negative charges that make it up can separate and redistribute themselves.)

    (b) Suppose the sphere is not neutral, but has on it a positive charge Q. How will this charge Q distribute itself when there are no other charges around?

    (c) Suppose the positive point charge q is brought near (but not in contact with) the sphere when it has on it a positive charge Q. Describe qualitatively how the charge will now distribute itself within and on the surface of the sphere.

[28] You certainly don't want to do a surface integration over the tetrahedron, and you don't have to: contemplate Gauss's law in a Zen kind of way. Or maybe a classical Greek sort of way, since they were so big on symmetry.

[29] What does Gauss's law tell you about the direction of the electric field in the neighborhood of the supposed equilibrium point and about the force that would therefore be experienced by a point charge at that point?

18. A conducting sphere carries a net charge q. Another conducting sphere, which has twice the radius, carries zero net charge.

    (a) If the two spheres are kept far apart but connected by a conducting wire so that charge can flow between them, how will the charge q divide itself between the two spheres? See the footnote if you need a hint [30]

    (b) What quantitative condition would correspond to the spheres being far apart?

    (c) Would the same distribution of charge result if the spheres were not far apart when they were connected by the wire, or if they were simply touched together rather than connected by a wire?

[30] Consider carefully what quantity is equalized when the two spheres are connected. If you cannot justify the equality you set up by a fundamental physical principle, you are wrong.

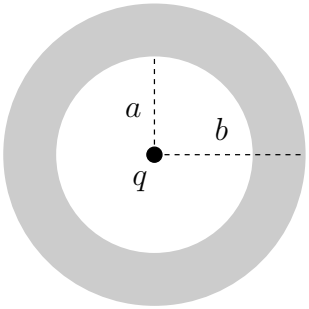


Figure 15.30: Problem 19

19. A positive point charge q is at the center of an uncharged spherical conducting shell of inner radius a and outer radius b, as shown in fig. (15.30). There are no other isolated charges in the vicinity.

    (a) What are the charges q_i and q_o, respectively, on the inner and outer surfaces of the shell?

    (b) Determine the electric field everywhere — inside, within, and outside the shell.

    (c) How would your results for q_i and q_o change if another positive point charge Q were present outside the shell?

    (d) Consider the forces that q, q_i, q_o, and Q exert on each other. Are the forces that q, the shell, and Q experience consistent with the Newton's third law, or was Newton just a total loser? Either way, explain — you can't get away with just glibly asserting that the third law is obeyed.

    (e) How would your answers to #19a through #19d be altered if the charge q were not at the center of the shell?

    (f) Suppose that the spherical conducting shell were instead a hollow conducting cube (that is, a conducting cube with a hollow concentric cubic interior). What, if anything, could you conclude about the electric field in the hollow interior to the cubic shell, within the body of the cubic shell itself, or outside of the cubic shell? What, if anything, could you conclude about the total charges and the distributions of charge on the inner and outer surfaces of the cubic shell?

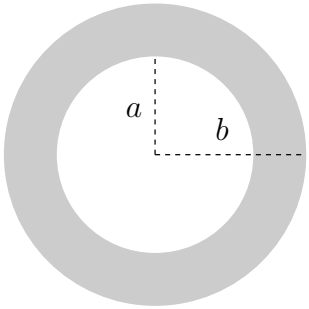


Figure 15.31: Problem 20

20. The spherical shell of charge shown in fig. (15.31) has inner radius a and outer radius b and is carries a uniform positive charge density ρ.

    (a) Determine the charge on the shell.

    (b) Determine the electric field everywhere (that is, at all distances r from the center of the shell: $r \leq a$, $a \leq r \leq b$, and $r \geq b$).

    (c) Determine the electrostatic potential everywhere, taking the potential to vanish out at infinity.

    (d) Determine the electrostatic potential everywhere, taking the potential to vanish at the center of the sphere. Think before you calculate.

    (e) Of course you obtained your results for the electrostatic potential from your results for the electric field, but if your results for the electrostatic potential were your starting point, would they be enough to determine all the components of the electric field by $\mathbf{E} = -\nabla\varphi$?

    (f) Determine the contribution of the shell to the electrostatic potential energy of a solid sphere of uniform charge density ρ' and radius c that lies outside the shell at a center-to-center distance R from it.

    (g) Explain why you would *not* get the same contribution if the solid sphere carried the same total charge but were instead a conducting sphere. Would the contribution be greater than or less than what you got in the preceding part?

    (h) Explain why the uniform charge distribution specified for the shell is unrealistic, in the sense that in the real world it is neither likely to occur nor feasible to produce.

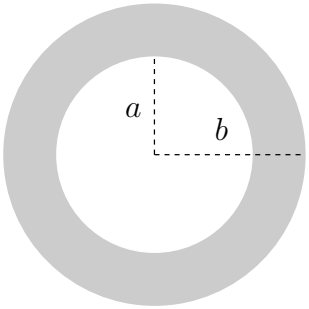


Figure 15.32: Problem 21

21. Fig. (15.32) now shows the cross section of a cylindrical shell of inner radius a and outer radius b that carries a uniform positive charge density λ per unit length. How's that for versatility? Anyway,

    (a) Determine the electric field everywhere (that is, at all distances r from the axis of the cylinder: $r \leq a$, $a \leq r \leq b$, and $r \geq b$).

    (b) Determine the electrostatic potential everywhere, taking the potential to vanish wherever strikes your fancy.

22. An infinite planar slab of finite thickness a is centered on the xy plane, so that it extends to $\pm\infty$ in the x and y directions and from $z = -\frac{1}{2}a$ to $z = +\frac{1}{2}a$. The slab carries a uniform charge density ρ per unit volume.

    Determine the electric field everywhere (that is, for $z \leq -\frac{1}{2}a$, $-\frac{1}{2}a \leq z \leq \frac{1}{2}a$, and $z \geq \frac{1}{2}a$).

23. Suppose the charge density within a solid sphere of radius a is given by

$$\rho = \rho_0 \frac{a}{r}$$

    where ρ_0 is a positive constant.

    (a) Determine the charge on the sphere.

    (b) Determine the electric field everywhere.

    (c) Determine the electrostatic potential everywhere, taking that potential to vanish out at infinity.

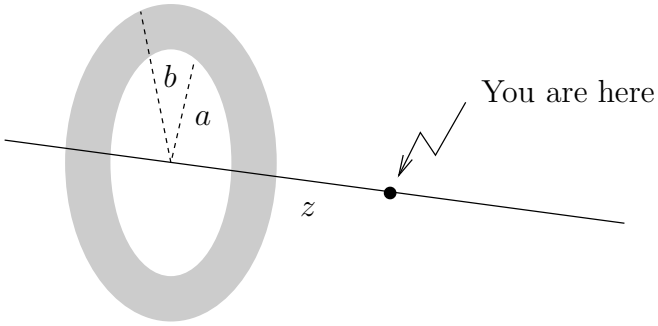


Figure 15.33: Problem 24

24. (a) A positive charge q is evenly spread over an annulus (that is, a washer) of inner radius a and outer radius b, as shown in fig. (15.33). By integrating over the charge distribution, show that the electric field at points along the axis of the annulus is given by

$$E = \frac{1}{4\pi\epsilon_0}\,\frac{2qz}{b^2-a^2}\left(\frac{1}{\sqrt{z^2+a^2}} - \frac{1}{\sqrt{z^2+b^2}}\right) \tag{15.27}$$

where z is the distance from the center of the annulus.

(b) Show that for large z the field of eq. (15.27) reproduces the electric field of a point charge.

(c) Show that when $a \to 0$ the field of eq. (15.27) reproduces our result (15.13) of p.790,

$$\frac{1}{4\pi\epsilon_0}\,\frac{2q}{R^2}\left(1 - \frac{z}{\sqrt{R^2+z^2}}\right)$$

for the electric field of a disk of charge of radius R.

(d) Show that when $a \to 0$ and $b \to \infty$ the field of eq. (15.27) reproduces the electric field of an infinite sheet of charge. See the footnote if you need a hint.[31]

(e) By integrating over the charge distribution, show that the electrostatic potential at points along the axis of the annulus is given by

$$\varphi = \frac{1}{4\pi\epsilon_0}\,\frac{2qz}{b^2-a^2}\left(\sqrt{z^2+b^2} - \sqrt{z^2+a^2}\right) \tag{15.28}$$

when the potential is taken to vanish out at infinity.

Continued on next page ...

[31] Since the total charge on an infinite sheet will be infinite, you want to rewrite eq. (15.27) in terms of the charge density rather than the charge q.

(f) Show that you obtain the same result (15.28) for the potential by integrating the electric field of eq. (15.27).

(g) Show that for large z the potential of eq. (15.28) reproduces the potential of a point charge.

(h) If a negative point charge $-Q$ of mass m is released from rest infinitely far out on the axis of the annulus, how fast will it be moving when it reaches a distance z from the center of the annulus?

(i) What is wrong with the preceding question?

25. Determine the electric field at the center of a spherical surface of radius a if one hemisphere of that surface carries a uniform positive charge density σ.

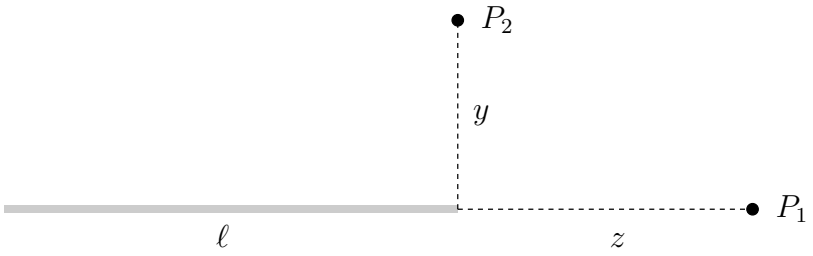


Figure 15.34: Problem 26

26. A thin rod of length ℓ carries a uniform linear charge density λ. As shown in fig. (15.34), the point P_1 is on the axis of the rod, a distance z from its end, and the point P_2 is a perpendicular distance y from, and aligned with, the end of the rod.

(a) Determine the electric field of the rod at point P_1.

(b) Show that your result for the preceding part reproduces the field of a point charge

i. For large z.

ii. As $\ell \to 0$.

(c) Determine the electrostatic potential at point P_1, taking the potential to vanish out at infinity.

(d) Show that your result for the preceding part reproduces the potential of a point charge

i. For large z.

ii. As $\ell \to 0$.

Recall that for $|u| \ll 1$, $\ln(1+u) \approx 1+u$.

Continued on next page ...

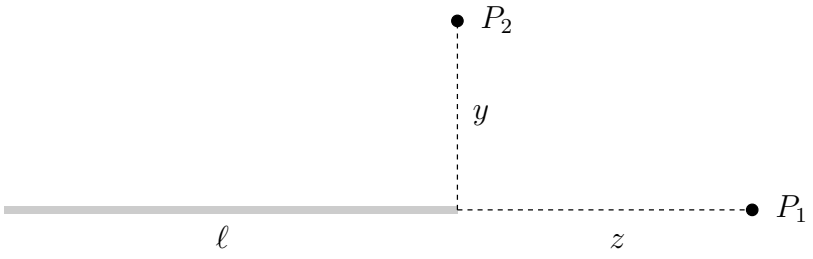


Figure 15.34: Problem 26

(e) Determine the parallel component of the electric field of the rod at point P_2. (*Parallel component* meaning, of course, the component parallel to the rod.)

(f) Determine the perpendicular component of the electric field of the rod at point P_2.

(g) Show that your results for the two preceding parts can be used to reproduce the field of an infinite line of charge.

(h) Determine the electrostatic potential at point P_2, taking the potential to vanish out at infinity.

For this purpose, you may find it helpful to note that

$$\int \frac{du}{u\sqrt{u^2+\ell^2}} = -\frac{1}{\ell}\,\sinh^{-1}\frac{\ell}{u} \qquad \int \frac{du}{\sqrt{u^2+\ell^2}} = \sinh^{-1}\frac{u}{\ell}$$

Then again, you might not. Who knows.

27. The inner of two thin coaxial cylindrical shells has radius a and carries a uniform positive linear charge density λ; the outer has radius b and carries equal and opposite charge density $-\lambda$.

(a) Show that magnitude of the potential difference between the shells is given by

$$|\Delta\varphi| = \frac{\lambda}{2\pi\epsilon_0}\,\ln\frac{b}{a}$$

(b) Which shell is at the higher potential?

28. A negative point charge $-q$ of mass m is placed at the center of a ring of radius a that carries a uniform positive charge density λ.

(a) Determine the angular frequency of small oscillations along the axis of the ring about this equilibrium point.

(b) Is this equilibrium stable or unstable in the plane of the ring?

29. A positive point charge q is placed in the plane of two infinite parallel lines of uniform positive charge density λ, halfway between them. The lines are a perpendicular distance ℓ apart.

    (a) Determine the angular frequency of small oscillations in the direction perpendicular to the lines.

    (b) Describe the motion that would result if q, at rest, were given a slight nudge that had a component parallel to the lines as well as a component perpendicular to them.

    (c) Describe the motion that would result if q, at rest, were given a slight nudge that had a component perpendicular to the plane of the lines as well as a component perpendicular to the lines themselves.

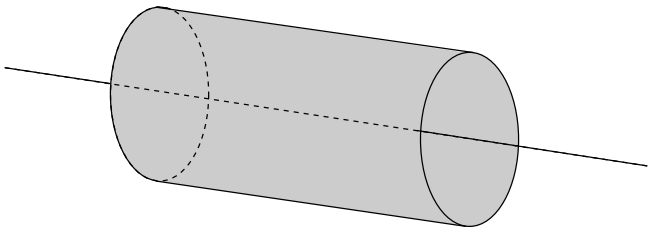


Figure 15.35: Problem 30

30. The thin cylindrical shell shown in fig. (15.35), of radius a and length ℓ, is coaxial with the z axis with its center at $z = 0$, so that it extends from $z = -\frac{1}{2}\ell$ to $z = +\frac{1}{2}\ell$. The shell carries a uniform surface charge density σ.

    Determine the electric field at locations on that part of the z axis that lies outside the shell. See the footnote if you need a hint.[32]

31. A point charge q is in the vicinity of an uncharged, thin spherical conducting shell. Make a rough sketch of the electric field and equipotential lines everywhere, including the region inside the shell.

    Note that *rough* does not mean *sloppy*, *lazy*, or *lame* — be sure to give a picture both full and free from heinous errors.

[32] You can integrate over a succession of rings of charge, for which we already have result (15.11) of p.788 for the field.

15.12 Sketchy Answers

(4) $\sqrt{\dfrac{2eV_0}{m_e}}$.

(5) $\dfrac{2msv_0^2}{e\ell^2}$.

(6) Magnitude is $\dfrac{mg\tan\theta}{E}$.

(7a) Magnitude is $4\pi\epsilon_0 V_0 a$.

(7b) $4\pi\epsilon_0\,\dfrac{V_0^2a^2}{R}$.

(8b) Magnitude is $\dfrac{1}{4\pi\epsilon_0}\,\dfrac{7q^2}{4a^2}$.

(8f) Magnitude is $\dfrac{1}{4\pi\epsilon_0}\,\dfrac{11q}{18a^2}$.

(8g) $-\dfrac{1}{4\pi\epsilon_0}\,\dfrac{7q^2}{2a}$.

(8h) $-\dfrac{1}{4\pi\epsilon_0}\,\dfrac{q}{3a}$.

(8j) Can't you read?

(9(a)i) Magnitude is $\dfrac{1}{4\pi\epsilon_0}\bigl(1+\sqrt{2}\bigr)\,\dfrac{qQ}{a^2}$.

(9b) $\alpha = -2\sqrt{2}$.

(10a) $x = -\frac{1}{3}\ell$, $Q = -\frac{16}{9}q$.

(11a) $x = -\frac{1}{3}\ell,\ \frac{1}{5}\ell$.

(11b) $x = -\ell$.

(11c) Didn't we warn you that we'd be asking some nonsense questions?

(11d) On the surfaces $15(x^2+y^2+z^2)+2x\ell-\ell^2=0$.

(12a) Magnitude is $\dfrac{1}{4\pi\epsilon_0}\,\dfrac{2qa}{(x^2+a^2)^{\frac{3}{2}}}$.

(12f) Magnitude is $\dfrac{1}{4\pi\epsilon_0}\,\dfrac{4qa|y|}{(y^2-a^2)^2}$.

(12h) $\pm\dfrac{1}{4\pi\epsilon_0}\,\dfrac{2qa}{y^2-a^2}$.

(13a) Magnitude is $\dfrac{1}{4\pi\epsilon_0}\,\dfrac{2q}{x^2}\left(1-\dfrac{x^3}{(x^2+a^2)^{\frac{3}{2}}}\right)$.

(13c) $-\dfrac{1}{4\pi\epsilon_0}\,\dfrac{2q}{x}\left(1-\dfrac{x}{\sqrt{x^2+a^2}}\right)$.

(13f) Magnitude is $\dfrac{1}{4\pi\epsilon_0}\,q\left(\dfrac{1}{(y-a)^2}+\dfrac{1}{(y+a)^2}-\dfrac{2}{y^2}\right)$.

(13h) $\pm\dfrac{1}{4\pi\epsilon_0}\,q\left(\dfrac{1}{y-a}+\dfrac{1}{y+a}-\dfrac{2}{y}\right)$.

(20a) $\frac{4\pi}{3}\rho(b^3-a^3)$.

(20b) You're on your own for the other regions, but within the shell it's

$$\frac{\rho r}{3\epsilon_0}\left(1-\frac{a^3}{r^3}\right)$$

(20c) You're on your own for the other regions, but within the shell it's

$$\frac{\rho}{3\epsilon_0}\left(\tfrac{3}{2}b^2-\tfrac{1}{2}r^2-\frac{a^3}{r}\right)$$

(20d) As you might have guessed, you're on your own for the other regions, but within the shell it's

$$\frac{\rho}{3\epsilon_0}\left(\tfrac{3}{2}a^2-\tfrac{1}{2}r^2-\frac{a^3}{r}\right)$$

(22) Within the slab you should get $\dfrac{\rho z}{\epsilon_0}$.

(23a) $2\pi\rho_0 a^3$.

(23b) You're on your own for outside the sphere, but inside it's $\dfrac{\rho_0 a}{2\epsilon_0}$.

(23c) You're on your own for outside the sphere, but inside it's $\dfrac{\rho_0}{2\epsilon_0}(2a-r)$.

(24h) $\sqrt{\dfrac{1}{4\pi\epsilon_0}\,\dfrac{4qQz}{m(b^2-a^2)}\left(\sqrt{z^2+b^2}-\sqrt{z^2+a^2}\right)}$.

(25) $\dfrac{\sigma}{4\epsilon_0}$.

(26a) $\dfrac{\lambda}{4\pi\epsilon_0}\left(\dfrac{1}{z}-\dfrac{1}{z+\ell}\right)$.

(26c) $\dfrac{\lambda}{4\pi\epsilon_0}\,\ln\left(\dfrac{z+\ell}{z}\right)$.

(26e) $\dfrac{\lambda}{4\pi\epsilon_0}\,\dfrac{1}{y}\left(1-\dfrac{y}{\sqrt{y^2+\ell^2}}\right)$.

(26f) $\dfrac{\lambda\ell}{4\pi\epsilon_0}\,\dfrac{1}{y\sqrt{y^2+\ell^2}}$.

(26h) $\dfrac{\lambda}{4\pi\epsilon_0}\,\sinh^{-1}\!\left(\dfrac{\ell}{y}\right)$.

(28a) $\sqrt{\dfrac{q\lambda}{2\epsilon_0 m a^2}}$.

(29a) $\dfrac{2}{\ell}\sqrt{\dfrac{q\lambda}{\pi\epsilon_0 m}}$.

(30) $\dfrac{\sigma a}{2\epsilon_0}\left(\dfrac{1}{\sqrt{a^2+\left(z-\frac{1}{2}\ell\right)^2}}-\dfrac{1}{\sqrt{a^2+\left(z+\frac{1}{2}\ell\right)^2}}\right)$.

Chapter 16

DC Circuits

16.1 Resistance & Power

As the term *circuit* would imply, electrical circuits involve the motion of charge around closed loops, and we will therefore be making use of our earlier definitions and results (14.3) through (14.6) for the various quantities related to the motion of charge. Recall from §14.2 that in terms of the charge density ρ per unit volume,

$$\rho = \frac{dq}{dV}$$

the current density $\mathbf{j}$ is

$$\mathbf{j} = \rho \mathbf{v} \tag{16.1}$$

where $\mathbf{v}$ is the velocity with which the charge is moving. This current density points in the direction of motion of the charge dq and has dimensions of charge per unit time per unit area: as we found in §14.2, to get the time rate

$$I = \frac{dq}{dt} \tag{16.2}$$

at which charge is flowing through a surface S, we integrate over that surface the component of $\mathbf{j}$ passing through it:

$$I = \int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{j} \tag{16.3}$$

This current I is thus specifically the time rate at which charge is passing *through a surface*.

In the context of an electrical circuit, it will be most useful to deal with the rate $I = dq/dt$ at which charge is flowing down wires or through circuit elements, that is, the rate at which charge is passing by the various points in the circuit. While we will seldom have occasion to refer to it explicitly, the corresponding area through which I is flowing is the cross-sectional area of the wire or circuit element.

Except for superconductors, which, for quantum-mechanical reasons, constitute perfect conductors, all substances, even good conductors like copper wire, present at least some resistance to the charge flowing through them. The effect is in some ways similar to velocity-dependent friction:[1] recall from §4.7 that for a mass m falling under the influence of gravity and a frictional force proportional to its velocity, the equation of motion

$$mg - \beta v = ma = m\frac{dv}{dt}$$

yielded the terminal velocity

$$v_{terminal} = \frac{mg}{\beta}$$

In the context of current flow, the gravitational force mg would become the electrical force qE, so that the terminal velocity of the electrons through a substance would, by analogy, be proportional to E. The current density $\mathbf{j} = \rho\mathbf{v}$ would then also be proportional to the electric field — not only in magnitude, but, since the electric force $\mathbf{F} = q\mathbf{E}$ on positive charges is in the same direction as the field, as a vector:

$$\mathbf{j} = \sigma\mathbf{E} \tag{16.4}$$

where the constant of proportionality σ is called the conductivity.[2] According to eq. (16.4), the current density at any given location within a substance is proportional to the electric field at that location, and although the arguments by which we have arrived at this conclusion were pretty hokey, it turns out that this is in fact true to a very good approximation for many substances as long as the electric field is of reasonable magnitude. In our system of units the values of the conductivity σ span a huge range, from zero for a vacuum and very nearly zero for poor conductors like glass, to about 10^{10} for the best metal conductors at low temperatures. For superconductors σ is infinite, corresponding to the vanishing of the electric field $\mathbf{E}$ within them.

The resistivity is defined to be the reciprocal of the conductivity and is, with another unfortunate collision in notation, conventionally denoted by ρ:

$$\rho = \frac{1}{\sigma} \tag{16.5}$$

[1] *In some ways similar to* indicates that we are making a very loose argument merely to establish plausibility. Even if we suppose that electrical resistance is analogous to mechanical friction, we have no justification for assuming that that friction is proportional to velocity. The real basis for electrical conductivity and resistance is quite involved and considerably beyond our scope.

[2] And is, of course, not to be confused with the σ used to denote surface charge density.

In most circuit applications, the electric field is constant along the length ℓ of a substance through which current is flowing, so that the voltage difference V along the substance is[3]

$$V = \left| -\int \mathbf{E} \cdot d\mathbf{r} \right| = E\ell \tag{16.6}$$

The current density $\mathbf{j}$ is therefore uniform over the cross-sectional area A of the substance and perpendicular to it, so that

$$I = \int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{j} = jA \tag{16.7}$$

Using eqq. (16.5), (16.6), and (16.7), we can rewrite eq. (16.4) as

$$\frac{I}{A} = \frac{1}{\rho}\frac{V}{\ell}$$

which, when solved for V, gives us Ohm's bogus law

$$V = IR \tag{16.8}$$

where the resistance R is defined to be

$$R = \frac{\rho\ell}{A} \tag{16.9}$$

In contrast to Coulomb's semibogus law, which is perfectly valid as long as certain restrictions are met, Ohm's bogus law derives — and rather imperfectly, at that — from $\mathbf{j} = \sigma\mathbf{E}$, which, while often a good approximation, is still only an approximation, and in fact a miserable one for many substances and conditions. So what most people refer to as Ohm's law would more rightly be called *Ohm's bogus law* or *Ohm's pretty good approximation* or perhaps *Ohm's not really a law*. But, hey, we don't make the rules.[4]

Anyway, a circuit element *designed* to have a substantial electrical resistance is called a resistor. Recall that our unit of current was the amp: $1\,\mathrm{A} = 1\,\mathrm{C/sec}$. Our unit of voltage is the volt (V), and that of resistance is the Ohm (Ω): $1\,\mathrm{A} \times 1\,\Omega = 1\,\mathrm{V}$. In actual circuits, voltage differences are applied across resistors by connecting them to voltage sources like batteries. Ohm's bogus law (16.8) tells us that, as an inverse proportion, the greater the resistance of a resistor, the less current flows through it in response to a given applied voltage.

[3] Most of the time, we use the big-people symbol φ for the electrostatic potential (a.k.a. voltage), but in the context of electrical circuits virtually everyone uses V (or, for reasons we will get to in Chapter 18, a fancy-shmancy curly E that looks like $\mathcal{E}$).

[4] While we're on the subject, and venting spleen, we might also note having once seen an engineering textbook that presented "the three forms of Ohm's law": $V = IR$, $I = V/R$, and $R = V/I$. Woof.

On the atomic level, electrical resistance originates in the collisions that take place between the electrons whose motion constitutes the current and the molecules of the substance through which those electrons are passing. In these collisions, the electrons lose some of their kinetic energy to the molecules, with the result that the random thermal motion of the molecules increases: the substance through which the current is flowing heats up. This is why you can use electricity to burn things up. Recall now the relation (15.23) between potential energy U and the potential (voltage) φ:

$$U = q\varphi$$

If there is a drop V in voltage as charge dq crosses a circuit element, the corresponding loss of energy is therefore

$$dU = dq\,V$$

which, if we divide both sides by dt and use eq. (16.2), becomes

$$\frac{dU}{dt} = \frac{dq}{dt}\,V = IV$$

On the left-hand side, the rate of loss of energy dU/dt is by definition the power P, so that we arrive at

$$P = IV \tag{16.10}$$

for the power dissipated in a circuit element through which a current I is flowing and across which there is a voltage drop V. Note that our derivation of this relation was quite general; eq. (16.10) applies to any kind of circuit element, not just to resistors. When we are dealing with resistors, Ohm's bogus law $V = IR$ allows us to re-express eq. (16.10) in two other, equivalent forms that are often convenient:

$$P = IV = I^2R = \frac{V^2}{R} \tag{16.11}$$

16.2 Series & Parallel Connections

Fig. (16.1) shows the conventional symbols used in circuit diagrams for DC voltage sources (such as batteries) and resistors. At least, one variation on

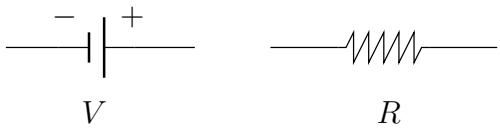


Figure 16.1: Symbols for Batteries and Resistors

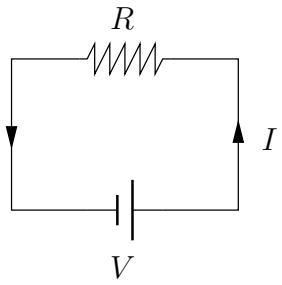


Figure 16.2: The Simplest Possible Circuit

the conventional symbols; the ± signs on the voltage source are (as will in fact be our own wont) often omitted, and sometimes there are two or more alternating pairs of short and long lines. But everyone adheres to the convention that the "short" and "long" sides of the voltage source represent, respectively, its lower- and higher-voltage terminals — that is, the negative and positive terminals. A resistor, in contrast, is symmetric — it doesn't matter which way current goes through it — , so it has a symmetric symbol — one that should lie comfortably within everyone's artistic ability to reproduce reliably.

In fig. (16.1) the lines emanating from the left and right sides of both the voltage source and the resistor represent the wires connected to them. The simplest possible circuit, shown in fig. (16.2), would consist of wires connecting the terminals of a single battery across a single resistor. From eq. (15.23),

$$\Delta U = q\Delta\varphi$$

we see that, for a positive charge, going "downhill" in potential energy U means going from higher to lower potential φ. Current will therefore flow from the positive to the negative terminal of the battery by traveling counterclockwise around the circuit and through the resistor.[5] For this circuit, the voltage across the resistor is just the voltage of the battery, and the current through the resistor, the current traveling around the circuit, and the current supplied by the battery are all the same current. So Ohm's bogus

[5]This is, of course, a lie: it is actually the negatively charged electrons, rather than the positively charged nuclei, that move through circuits. And by $\Delta U = q\Delta\varphi$, negative charges go downhill in potential energy by flowing in the opposite direction: out of the negative terminal of the battery, clockwise around the circuit, and back into the positive terminal of the battery. Mathematically, however, the flow of negative charge is equivalent to the flow of positive charge in the opposite direction, and even though the reality is otherwise, it is conventional in circuit applications to think in terms of the flow of positive charge. Ordinarily, there is nothing so pernicious as humanity's tendency to believe what it wants to believe in spite of the facts, but in this case no harm is done.

law for this circuit can be applied as

$$V_{battery} = I_{around\ circuit}\, R$$

The power supplied by the battery to the circuit is

$$P_{battery} = I_{from\ battery} V_{battery}$$

which is, as required by conservation of energy, the same as the power dissipated in the resistor:

$$P_{resistor} = I_{through\ resistor} V_{across\ resistor}$$

If this all seems pretty tautological, that's because there's only one voltage, one current, and one resistor in this circuit — hardly fertile ground for profound observations. Real circuits involve more than just a single battery hooked up across a single resistor, and the two most basic ways to connect circuit elements are in series and in parallel. While there are more complicated ways to connect circuit elements — ways that must be analyzed by different methods that we'll discuss later in the chapter and apply in some of the problems — , most circuits are simply combinations of series and parallel connections.

Figure 16.3: Resistors Connected in Series

As the term would imply, a series connection is a concatenation of circuit elements like that shown in fig. (16.3). When resistors are connected in series, there is only one path for the current to take, so the same current I must flow through each of them.[6] The voltage $V_{applied}$ applied across the series combination, on the other hand, will be divided among the individual resistors, with the sum of the voltage drops across each resistor equaling the applied voltage:

$$V_{applied} = \sum_{\text{resistors}} V_{resistor}$$

Since the same current I flows through each resistor, Ohm's bogus law gives

$$V_{resistor} = IR_{resistor}$$

for each of them, so that we have

$$V_{applied} = \sum_{\text{resistors}} IR_{resistor}$$

[6] If any current were to just stop at some point along the way, there would be an accumulation of charge at that point. Because this charge would repel itself, such an accumulation will not occur, and the current will keep moving.

$$= I \sum_{\text{resistors}} R_{resistor}$$

We can therefore apply Ohm's bogus law directly to the series combination of resistors in the form

$$V_{applied} = I R_{eq,\ series}$$

if we take

$$R_{eq,\ series} = \sum_{\text{resistors}} R_{resistor} \tag{16.12}$$

as the equivalent resistance of the resistors connected in series. In other words, a series combination of resistors behaves in a circuit as though, instead of a set of resistors in series, you had a single resistor of resistance $R_{eq,\ series}$. We'll work through an example to illustrate how this equivalent resistance can be used to analyze a circuit at the end of this section. But first:

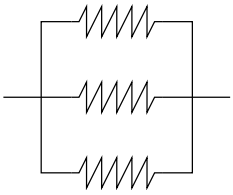


Figure 16.4: Resistors Connected in Parallel

A parallel connection of resistors is such that the applied voltage is applied in full across each resistor individually, as shown for the case of three resistors in fig. (16.4). Note that the "pitchfork" of wires on the left ensures that the left side of each of the resistors is at the same voltage: where the wires of the pitchfork are connected together, they become effectively a single, continuous conductor, all of which is, by virtue of being a conductor, at the same voltage. Whatever voltage is applied to the "handle" of the pitchfork is thus conveyed down each "tine" and applied to the left side of each of the resistors. Likewise, the right pitchfork ensures that the right side of each of the resistors is at the same voltage. The voltage difference across each of the resistors thus matches the voltage applied to the combination. Such connections are called *parallel* because in circuit diagrams the resistors (or whatever circuit elements are being connected in parallel) are conventionally drawn geometrically parallel to each other, as in fig. (16.4).

Since the voltage across each of the resistors is the same as the voltage $V_{applied}$ applied to the parallel combination, Ohm's bogus law gives for each resistor

$$V_{applied} = I_{resistor}\ R_{resistor}$$

and hence

$$I_{resistor} = \frac{V_{applied}}{R_{resistor}}$$

For the series combination, the same current flowed through each resistor but the voltages differed; for the parallel combination, it is the voltages that are the same and the currents through the resistors that differ: the current I_{total} that flows into the combination along the stem of the pitchfork will, when it reaches the branching point, split up into the currents that flow through each of the individual resistors. The sum of the currents through the individual resistors should therefore equal the total current:

$$\begin{aligned} I_{total} &= \sum_{\text{resistors}} I_{resistor} \\ &= \sum_{\text{resistors}} \frac{V_{applied}}{R_{resistor}} \\ &= V_{applied} \sum_{\text{resistors}} \frac{1}{R_{resistor}} \end{aligned}$$

If we compare this to Ohm's bogus law $V = IR$ written in the form

$$I_{total} = V_{applied} \, \frac{1}{R_{eq,\ parallel}}$$

we see that the parallel combination of resistors is equivalent to the single resistance

$$\frac{1}{R_{eq,\ parallel}} = \sum_{\text{resistors}} \frac{1}{R_{resistor}} \tag{16.13}$$

Circuits will generally consist, not of purely series or parallel combinations of resistors, but of nested series and parallel combinations. To collapse these circuits down to a single equivalent resistance, you work from the inside out, looking for combinations within the circuit that are purely series or parallel, collapsing these down to an equivalent resistance, then repeating the process with what remains of the circuit. Consider, for example, the combination of four resistors shown at the top of fig. (16.5). For simplicity, let us take the values of the resistors to be

$$R_1 = 3\,\Omega \qquad R_2 = 6\,\Omega \qquad R_3 = 6\,\Omega \qquad R_4 = 4\,\Omega$$

To break this combination down, we first note that R_2 and R_3 are connected in series, so that together they are equivalent to a single resistance

$$R_{23} = R_2 + R_3 = 6 + 6 = 12\,\Omega$$

The combination that remains after this pair has been collapsed is shown in the middle of fig. (16.5). Now we can collapse the parallel combination of R_{23} and R_4 down to

$$\frac{1}{R_{234}} = \frac{1}{R_{23}} + \frac{1}{R_4} = \frac{1}{12} + \frac{1}{4} = \frac{1}{3} \qquad\qquad R_{234} = 3\,\Omega$$

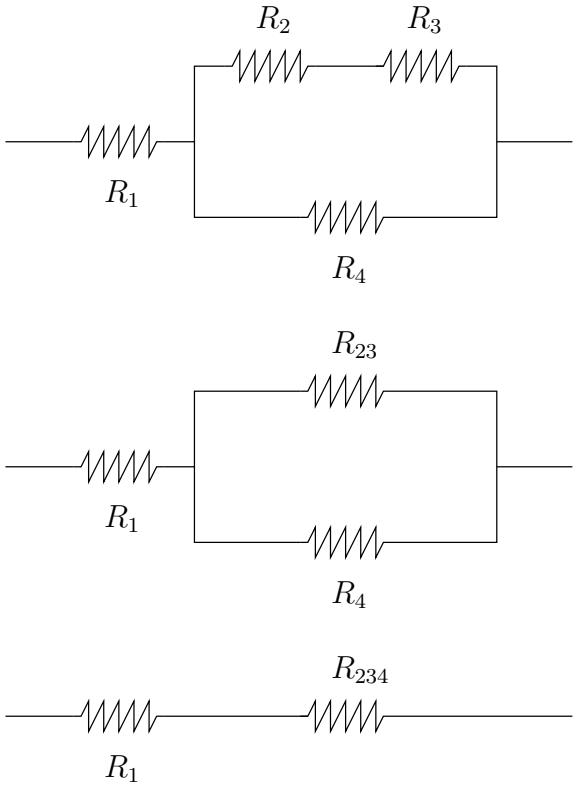


Figure 16.5: "Fun" and Games

We are then left with the simple series combination shown at the bottom of fig. (16.5), so that the equivalent resistance we arrive at for the combination of all four resistors is

$$R_{1234} = R_1 + R_{234} = 3 + 3 = 6\,\Omega$$

If this combination of four resistors were hooked up across a 12 V battery, together they would behave as a single 6 Ω resistor, and by Ohm's bogus law the current flowing from the battery through the combination would therefore be

$$V = IR \qquad 12 = I(6) \qquad I = 2\,\text{A}$$

All of this current would pass through R_1, and applying Ohm's bogus law to R_1 gives

$$V = IR = 2(3) = 6\,\text{V}$$

for the voltage across R_1. Since 6 of the 12 applied volts are dropped across R_1, the remainder of the combination (what we have called R_{234}) must have

$12 - 6 = 6\,\text{V}$ across it.[7] Since R_{23} and R_4 are in parallel, this means that there are 6 V across R_{23} and 6 V across R_4. Ohm's bogus law then gives for the current through these resistances[8]

$$V_{23} = I_{23}R_{23} \qquad 6 = I_{23}(12) \qquad I_{23} = \tfrac{1}{2}\,\text{A}$$
$$V_4 = I_4 R_4 \qquad 6 = I_4(4) \qquad I_4 = \tfrac{3}{2}\,\text{A}$$

And since the same $\frac{1}{2}$ A must flow through both R_3 and R_2, Ohm's bogus law for these resistors gives[9]

$$V_2 = I_2 R_2 = \tfrac{1}{2}(6) = 3\,\text{V} \qquad V_3 = I_3 R_3 = \tfrac{1}{2}(6) = 3\,\text{V}$$

In this last step, we could also have made use of symmetry: because $R_2 = R_3$, the 6 V across this series combination must be divided equally between the two resistors, 3 V each.[10] Making use of such symmetries within a circuit will often greatly simplify its analysis. If, for example, the three resistances connected in parallel in fig. (16.6) are equal to each other, then the current I that flows through the combination must split equally among them, $\frac{1}{3}I$ through each.

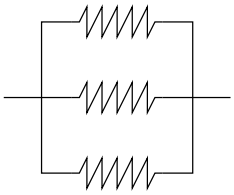


Figure 16.6: No Need for a Caption Here

Finally, note that whereas we worked from the inside out to determine the equivalent resistance of the combination of the four resistors in fig. (16.5), to determine the voltages and currents within the circuit we then worked from the outside in, applying Ohm's bogus law over and over again, *ad nauseam* — first to the circuit as a whole, then to various sets of resistors, and ultimately to individual resistors.

[7] We could also have obtained this result by applying Ohm's bogus law to R_{234}: the total current flowing through this combination of three resistors is the same as the 2 A flowing through R_1, so that

$$V_{234} = IR_{234} = 2(3) = 6\,\text{V}$$

There are usually many different ways to work a given circuit.

[8] Alternatively, once we know that $\frac{1}{2}$ A of the total of 2 A flows through R_{23}, we know that the remaining $\frac{3}{2}$ A must flow through R_4.

[9] And yet more redundancy: $3 + 3 = 6\,\text{V}$ ✓

[10] We could also have made use of symmetry in the series combination of R_1 and R_{234}, both of which were 3 Ω, to conclude at once that half of the applied 12 V must be across each of these two resistances.

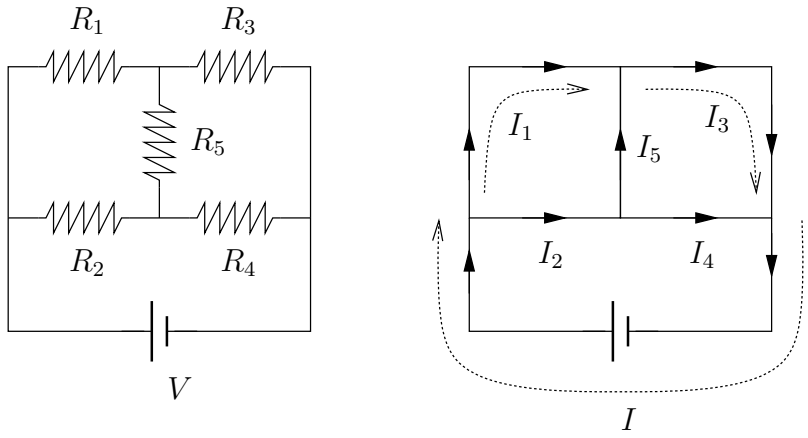


Figure 16.7: A Neither Series nor Parallel Connection

16.3 Loop & Junction Rules

Another technique for analyzing circuits is the method of loop and junction rules, sometimes also called Kirchhoff's rules. While this method is more cumbersome algebraically than that of equivalent resistance, it is more powerful and can be used to analyze circuits that cannot be resolved into combinations of series and parallel connections. The ideas behind both rules are very simple: The loop rule states that since returning to the same point means returning to the same voltage, the sum of the changes in voltage around any closed loop in a circuit must vanish. The junction rule states that since charge will not accumulate at any point in the circuit, the sum of the currents flowing into each junction of wires in the circuit must equal the sum of the currents flowing out of it.

Consider, for example, the circuit shown on the left side of fig. (16.7). Your first step in analyzing a circuit by the loop and junction rules will be to label all the currents in the circuit, as we have done on the right side of the figure. Although the current in each branch of the circuit is the same throughout that branch, every time a junction is reached, the current will split up into different currents. Thus we have noted in fig. (16.7) that the current I coming out of the battery flows from the positive terminal around to the junction on the left side of the circuit, where it will break up into currents that we have labeled I_1 and I_2. It is generally best to label currents on a separate outline of the circuit rather than on the circuit itself, which would become too cluttered. As we will see below, it is important to indicate the direction of each current. If you aren't sure about the direction of a current beforehand, just arbitrarily assign a direction to it; ultimately your solution for that current will tell you the direction in which it actually flows: if your solution is positive, that current really does flow the way you indicated; if your solution is negative, it flows the opposite way. There is no wrong choice

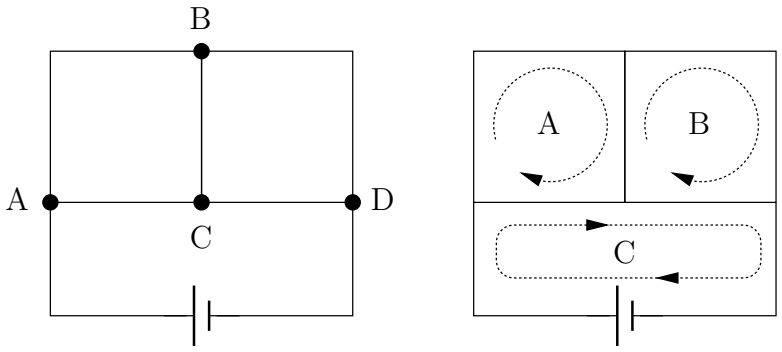


Figure 16.8: The Junctions and Loops

when you are labeling the directions of the currents: you will have plenty of opportunities to make mistakes when working with loop and junction rules, but your choice of directions for your currents is not among them.

The next step is of course to apply the loop and junction rules. For the junctions A, B, C, and D shown on the left side of fig. (16.8), the junction rule gives, respectively,

$$I = I_1 + I_2 \tag{16.14a}$$
$$I_1 + I_5 = I_3 \tag{16.14b}$$
$$I_2 = I_5 + I_4 \tag{16.14c}$$
$$I_3 + I_4 = I \tag{16.14d}$$

And if we go around loops A, B, and C as shown on the right side of fig. (16.8), the loop rule gives

$$-I_1R_1 + I_5R_5 + I_2R_2 = 0 \tag{16.15a}$$
$$-I_5R_5 - I_3R_3 + I_4R_4 = 0 \tag{16.15b}$$
$$V - I_2R_2 - I_4R_4 = 0 \tag{16.15c}$$

Note that some terms in the loop equations are positive and some negative: when we cross a resistor *with* a current (that is, in the same direction that we have labeled for it), the voltage *drops* by IR, so that the voltage change is $-IR$. Similarly, when we cross a resistor *against* the current, the voltage change is $+IR$. And when we cross a battery in what you might call the "forward" direction, that is, going from the negative to the positive terminal, we are going from lower to higher voltage, so that the change is $+V$; crossing a battery in the "backward" direction, going from the positive to the negative terminal, the change is $-V$.

The astute reader will have noticed that although we have only six unknown currents, $\{I, I_1, \ldots, I_5\}$, we have seven equations. In fact, while we have made exhaustive use of the four junctions in the circuit, we could, if we

were perverse, generate infinitely many more equations using the loop rule. If, for example, we went around the outer perimeter of the circuit clockwise, we would have

$$V - I_1 R_1 - I_3 R_3 = 0 \tag{16.16}$$

We could then go around the same loop, but taking a detour to loop twice around loop A. And so on. All of these equations would be correct and consistent, but they are of course redundant. Going clockwise around the outer perimeter is, in a geometric sense, equivalent to the sum of going clockwise around loops A, B, and C, and you can see that algebraically the sum of eqq. (16.15a), (16.15b), and (16.15c) does in fact yield eq. (16.16). Likewise only three of the four junction equations are independent. The general rule of thumb is to first generate a loop equation for each *distinct* (that is, geometrically independent) loop in the circuit, then generate as many equations by the junction rule as you need to solve for the unknowns. This does not by itself guarantee that you won't pick redundant junctions, but after a while you'll get a feel for the whole business. It's a Zen kind of thing.

Anyway, for this example, one set of independent equations is (16.14a) through (16.14c) together with (16.15a) through (16.15c). Solving this system of six linear equations in six unknowns, while a pain, would be straightforward. You just turn the crank.[11]

[11] In case you're burning with curiosity,

$$I = -\frac{((R_4 - R_3 - R_2 - R_1)\ R_5 + (R_2 + R_1)\ R_4 + (-R_2 - R_1)\ R_3)\ V}{((R_3 + R_1)\ R_4 + R_2 R_3 + R_1 R_2)\ R_5 + ((R_2 + R_1)\ R_3 - R_1 R_2)\ R_4 + R_1 R_2 R_3}$$

$$I_1 = -\frac{((R_4 - R_2)\ R_5 + R_2 R_4 - R_2 R_3)\ V}{((R_3 + R_1)\ R_4 + R_2 R_3 + R_1 R_2)\ R_5 + ((R_2 + R_1)\ R_3 - R_1 R_2)\ R_4 + R_1 R_2 R_3}$$

$$I_2 = \frac{((R_3 + R_1)\ R_5 - R_1 R_4 + R_1 R_3)\ V}{((R_3 + R_1)\ R_4 + R_2 R_3 + R_1 R_2)\ R_5 + ((R_2 + R_1)\ R_3 - R_1 R_2)\ R_4 + R_1 R_2 R_3}$$

$$I_3 = -\frac{((R_4 - R_2)\ R_5 + (R_2 + R_1)\ R_4)\ V}{((R_3 + R_1)\ R_4 + R_2 R_3 + R_1 R_2)\ R_5 + ((R_2 + R_1)\ R_3 - R_1 R_2)\ R_4 + R_1 R_2 R_3}$$

$$I_4 = \frac{((R_3 + R_1)\ R_5 + (R_2 + R_1)\ R_3)\ V}{((R_3 + R_1)\ R_4 + R_2 R_3 + R_1 R_2)\ R_5 + ((R_2 + R_1)\ R_3 - R_1 R_2)\ R_4 + R_1 R_2 R_3}$$

$$I_5 = -\frac{(R_1 R_4 + R_2 R_3)\ V}{((R_3 + R_1)\ R_4 + R_2 R_3 + R_1 R_2)\ R_5 + ((R_2 + R_1)\ R_3 - R_1 R_2)\ R_4 + R_1 R_2 R_3}$$

16.4 Capacitance

Just so we can get it out of the way and then get on with things: there is no such thing as a flux capacitor, not any more than there is a Santa Claus or an Easter Bunny.[12] And real capacitors have nothing to do with time travel. Sorry.

Though we won't prove it here, it turns out that for any configuration of conductors i, $i = 1, 2, 3, \ldots, n$, there is a linear relation between the voltages V_i on the conductors and the net charges q_i they carry:

$$q_i = \sum_{j=1}^{n} C_{ij} V_j$$

where the C_{ij}, known as the coefficients of capacitance, are constant and depend only on the geometries of the individual conductors and of their arrangement. For a pair of conductors, this reduces to simply

$$q = CV \tag{16.17}$$

where V is the voltage difference between the conductors, of which one carries charge q and the other charge $-q$.[13] We can use this relation to determine the capacitance C from the geometry of the conductors: the trick is to find, by some means, a relation between q and V and write it in the form $q =$ something $\times V$; the capacitance is then whatever sits in front of the V.

As with resistance and resistors, *capacitance* is the abstract property, and a capacitor is a circuit element designed to have a substantial capacitance. The question of why such a circuit element is useful we will postpone; for the moment we will merely note that our unit of capacitance is the Farad (F), short for Faraday: $1\,\mathrm{F} = 1\,\mathrm{C} \times 1\,\mathrm{V}$. Just as a Coulomb is a huge quantity of charge, a Farad is a huge capacitance; most capacitors used in circuits have capacitances measured in μF or even pF.

16.4.1 Parallel-Plate Capacitors

Consider first the case of two identical parallel conducting plates of area A, aligned flush with each other and close together (in the sense that the separation d between plates is very small compared to the size of the plates), with one plate carrying charge q and the other charge $-q$. As long as we are not close to the edges of the plates, the electric field between them will to a good approximation be the same as that of infinite plates, which we found in §15.1.4 to be $E = \sigma/\epsilon_0$, where $\sigma = q/A$ is the magnitude of the surface

[12] Our sincerest apologies if this leaves you reeling from a triple blow.

[13] The charges are equal and opposite because, as it turns out, this is the distribution of charge that minimizes energy.

charge density on the plates. Since this field is constant, the line integral $|-\int \mathbf{E}\cdot d\mathbf{r}|$ for the voltage difference V between the plates will yield simply $V = Ed$. Putting this all together, we have

$$V = Ed = \frac{\sigma}{\epsilon_0}\, d = \frac{q}{\epsilon_0 A}\, d$$

or, rewriting this in the form $q = \text{something} \times V$,

$$q = \frac{\epsilon_0 A}{d}\, V$$

We therefore conclude that the capacitance of parallel plates is

$$C = \frac{\epsilon_0 A}{d} \tag{16.18}$$

Note that our result for the capacitance does indeed depend only on the geometry of the plates (their area A and separation d).

Also be aware that this result is only an approximation for plates of finite area: near the edges of the plates the field and charge density will not have the constant values we have assumed above. In practice, however, the approximation is a very good one as long as edge effects are minimized by making the area of the plates very large and their separation very small.

16.4.2 Cylindrical Capacitors

Next consider the case of two coaxial cylindrical conducting shells of length ℓ, an inner shell of radius a and an outer shell of radius b, aligned flush with each other, with one shell carrying charge q and the other charge $-q$. As long as we are not close to the ends of the cylinders, the electric field between them will to a good approximation be the same as that given by eq. (15.4) for an infinite cylindrical shell: [14]

$$E = \frac{1}{4\pi\epsilon_0}\,\frac{2\lambda}{r}$$

where $\lambda = q/\ell$ is the magnitude of the linear charge density on the cylinders. The line integral for the voltage difference V between the shells thus yields

$$V = \left|-\int \mathbf{E}\cdot d\mathbf{r}\right| = \int_a^b \frac{1}{4\pi\epsilon_0}\,\frac{2\lambda}{r}\, dr = \frac{1}{4\pi\epsilon_0}\, 2\lambda\, \ln\frac{b}{a}$$

Putting this all together, we have

$$V = \frac{1}{4\pi\epsilon_0}\, 2\lambda\, \ln\frac{b}{a} = \frac{1}{4\pi\epsilon_0}\, 2\,\frac{q}{\ell}\, \ln\frac{b}{a}$$

[14] Recall that only the charge on the inner cylinder contributes to the electric field between them; the electric field of the outer shell is zero at points inside its radius.

or, solving for q in terms of V,

$$q = \frac{2\pi\epsilon_0\ell}{\ln(b/a)}\,V$$

We therefore conclude that the capacitance of a cylindrical capacitor is

$$C = \frac{2\pi\epsilon_0\ell}{\ln(b/a)} \tag{16.19}$$

This result, which again depends only on the geometry of the conductors (the parameters ℓ, a, and b), is only an approximation for finite cylinders, but a good one if they are long compared to their separation $b - a$.

When this separation $d = b - a$ between the cylinders is very small,

$$\ln\frac{b}{a} = \ln\frac{a+d}{a} = \ln\left(1 + \frac{d}{a}\right) \approx \frac{d}{a}$$

so that

$$C \approx \frac{2\pi\epsilon_0\ell}{d/a} = \frac{\epsilon_0(2\pi a)\ell}{d} = \frac{\epsilon_0 A}{d}$$

where $A = 2\pi a\ell$ is the area of each cylinder's side: as we would expect, when the cylinders are very close together their capacitance reproduces that of a parallel plates.

16.4.3 Spherical Capacitors

Finally, consider the case of two concentric spherical conducting shells, an inner shell of radius a and an outer shell of radius b, one shell carrying charge q and the other charge $-q$. The electric field between them is the same as that of a point charge,[15]

$$E = \frac{1}{4\pi\epsilon_0}\,\frac{q}{r^2}$$

The line integral for the voltage difference V between the shells therefore yields

$$V = \left|-\int \mathbf{E}\cdot d\mathbf{r}\right| = \left|-\int_a^b \frac{1}{4\pi\epsilon_0}\,\frac{q}{r^2}\,dr\right| = \frac{1}{4\pi\epsilon_0}\,q\left(\frac{1}{a} - \frac{1}{b}\right) = \frac{1}{4\pi\epsilon_0}\,q\,\frac{b-a}{ab}$$

and hence

$$q = 4\pi\epsilon_0\,\frac{ab}{b-a}\,V$$

[15] Again, recall that only the charge on the inner shell contributes to the electric field between them; the electric field of the outer shell is zero at points inside its radius.

We therefore conclude that the capacitance of concentric spherical shells is

$$C = 4\pi\epsilon_0 \frac{ab}{b-a} \tag{16.20}$$

As always, this result depends only on the geometry of the conductors (the radii a and b). This time, however, the result is exact: there are no edge effects.

As we would expect, our result for the capacitance of spherical shells reproduces that of parallel plates when the distance $d = b - a$ between the shells is very small:

$$\frac{ab}{b-a} = \frac{ab}{d} \approx \frac{R^2}{d}$$

where R is the common radius $R = a = b$ of the shells in the limit $a \to b$. Using this in eq. (16.20), we do in fact obtain

$$C \approx 4\pi\epsilon_0 \frac{R^2}{d} = \frac{\epsilon_0(4\pi R^2)}{d} = \frac{\epsilon_0 A}{d}$$

where $A = 4\pi R^2$ is the area of the shells.

A lone spherical shell is essentially a concentric pair, with the outer shell having infinite radius. If we let $b \to \infty$, eq. (16.20) gives

$$C = 4\pi\epsilon_0 \frac{ab}{b-a} = 4\pi\epsilon_0\, a \frac{b}{b-a} \to 4\pi\epsilon_0\, a$$

In this case, $q = CV$ is just restating the result for the voltage at the surface of a spherical shell of radius a, since

$$q = CV = 4\pi\epsilon_0\, aV$$

is equivalent to

$$V = \frac{1}{4\pi\epsilon_0} \frac{q}{a}$$

You should be able to work out the capacitances of the following geometries of conductors:

- Parallel plates.
- Coaxial cylindrical shells.
- Concentric spherical shells.
- A single spherical shell.

16.4.4 A Few Observations

Capacitance is so called because it is a measure of capacity for storing electric charge. Ordinarily the attraction between positive and negative charges, and also the repulsion of like charge for itself, will cause any isolated charge that you try to accumulate to disperse and neutralize itself. But by gathering opposite charges onto neighboring conductors, you are using the electrical attraction to advantage: it will hold the charge on the conductors. You need simply have a good insulator between the conductors to keep the charge from leaking from one conductor to the other.

As we will see in §16.5.1 and, more rigorously, §19.3, the electric field has an energy density (energy per unit volume) associated with it. It turns out that when this energy density is integrated over all of space, the total energy is lowest when the charges on the conductors are equal and opposite. In the case of a parallel-plate capacitor, for example, $E = 0$ outside the plates only when the charges on the plates are equal and opposite, and likewise for a cylindrical or spherical capacitor. In practice, you charge capacitors by hooking them up to batteries or other voltage sources: when the battery's voltage is applied across the conductors, electrons run out of the negative terminal onto what thus becomes the negatively charged plate, and an equal number of electrons leaves the other plate and flows back into the positive terminal of the battery, leaving that plate with a positive charge equal in magnitude to that of the negative plate. Held by their electrical attraction for each other, these charges remain on the plates even after the battery is disconnected from the capacitor.

In practice, capacitors consist of spiral rolls of foil "plates" separated by an insulating substance and tend therefore to be cylindrical in shape. Using such rolls rather than just a pair of parallel plates or coaxial cylinders makes the dimensions of the capacitor much smaller in proportion to its capacitance. There is also a semibogus sort of capacitor known as an electrolytic capacitor that, as its name implies, works by separation of ions, similar to the mechanism of a battery.[16] This chemical trick makes it possible to build capacitors with huge capacitances yet reasonable size: while a 1 F parallel-plate capacitor with plates separated by 1 mm would be over 10 km on a side, a 1 F electrolytic capacitor will fit in the palm of your hand. But whereas true capacitors are symmetric — either side could be the positive and either the negative — , electrolytic capacitors are, like batteries, asymmetric, and it is critical that they be connected into circuits the right way around: if hooked into a circuit backward, an electrolytic capacitor can explode.[17]

[16] Which, we presume, you learned about in a previous life in chemistry. Either that, or you are out of luck, for we have little taste for such chemical-engineering trivia.

[17] We know what you're thinking. Naughty, naughty!

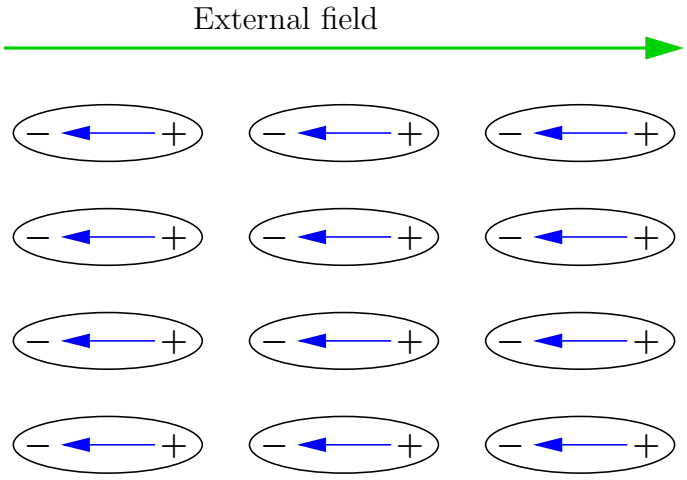


Figure 16.9: Dipoles of Dielectric Aligning with an External Field

16.4.5 Dielectrics

Dielectrics are substances whose molecules are polar, that is, electrically biased so that while neutral overall, they are positive at one end and negative at the other. In other words, the molecules of a dielectric substance are electric dipoles. If there is no externally applied electric field, thermal motion keeps the orientation of these dipoles random, so that their net electric field vanishes on a macroscopic level. But an external electric field will, as discussed in §15.5, exert a torque on the dipoles that will cause them to align with it and thus with each other, as shown to a very unrealistic degree in fig. (16.9): an external field to the right will push the positive ends of the molecules to the right and pull their negative ends to the left. As a result, the electric fields of the dipoles themselves (the blue arrows in fig. (16.9)) partially cancel the external field. If a dielectric is used between the plates of a capacitor, the effect will be to decrease the voltage difference V between the plates: since the net electric field is reduced by the backward contribution of the dipoles, the line integral $|-\int \mathbf{E}\cdot d\mathbf{r}|$ that gives the voltage difference between the plates will be reduced. Because this is accomplished without any change in the charge q on each plate, the capacitance C must increase to compensate for the decrease in V in $q = CV$.

While the expression for capacitance varies with the geometry of the capacitor, every capacitance has, as we have seen, a factor of ϵ_0. Ultimately this factor comes from Gauss's law,

$$\nabla\cdot\mathbf{E} = \frac{1}{\epsilon_0}\rho$$

The charge on the plates is the integral of the charge density ρ on the right-

hand side, and the voltage difference between the plates of the capacitor is the line integral of the electric field on the left-hand side. The relation of q to V will therefore always be of the form

$$V \sim \frac{1}{\epsilon_0} q$$

so that

$$q \sim \epsilon_0 V$$

and hence

$$C \sim \epsilon_0$$

We can therefore take into account the effect of the dielectric on capacitance by setting

$$\epsilon_0 \to \epsilon$$

where the electric permittivity ϵ of the dielectric substance differs from the electric permittivity ϵ_0 of the vacuum by a factor known as the dielectric's dielectric constant κ:

$$\epsilon = \kappa \epsilon_0$$

The effect of putting a dielectric between the plates of a capacitor is therefore

$$C \to \kappa C$$

For a vacuum, $\epsilon = \epsilon_0$ and hence $\kappa = 1$; for all other substances, $\kappa \geq 1$, though usually not much greater.

This reasoning we have used above might seem kind of bogus, and it is: the effect of dielectrics is much more general than their effect on capacitance, but we don't want to get that involved in what for us is only a rather minor aside. If we were to look into dielectrics more generally and rigorously, it would, however, turn out that they can in fact be taken entirely into account by a simple shift in the value of the electric permittivity and that our conclusions above are quite correct.

And that's about all we'll be saying about dielectrics.

16.5 Capacitors in Circuits

In circuit diagrams, capacitors are represented by what is supposed to look like parallel plates: a pair of transverse parallel lines, similar to the symbol for a battery, but with both sides of equal length. Fig. (16.10) shows a capacitor connected across a battery: conceptually, a positive charge $+q$ flows — essentially instantaneously — out of the positive terminal of the battery onto one plate, and the other plate acquires charge $-q$ as an equal charge

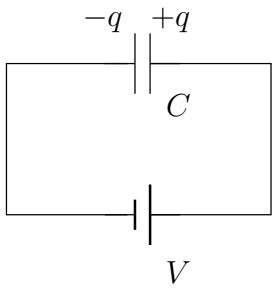


Figure 16.10: A Capacitor Connected to a Battery. Whoa!

$+q$ leaves it and flows into the negative terminal of the battery. Although no charge crosses the plates of the capacitor, in effect it is as though a charge $+q$ flows out of the positive terminal of the battery, through the capacitor, and back into the negative terminal of the battery.

As is the case for any circuit elements, when capacitors are connected in parallel the applied voltage $V_{applied}$ is applied fully across each capacitor individually. In response to this applied voltage, charge will accumulate on the plates of each capacitor: $+q_{capacitor}$ on one side, $-q_{capacitor}$ on the other, with

$$q_{capacitor} = C_{capacitor} V_{applied}$$

The total charge accumulating on the parallel combination will be simply the sum of these individual charges:

$$\begin{aligned} q_{total} &= \sum_{\text{capacitors}} q_{capacitor} \\ &= \sum_{\text{capacitors}} C_{capacitor} V_{applied} \\ &= \left(\sum_{\text{capacitors}} C_{capacitor} \right) V_{applied} \end{aligned}$$

If we were to apply $q = CV$ to the parallel combination in terms of an equivalent capacitance C_{eq}, we would write

$$q_{total} = C_{eq} V$$

We therefore conclude that when capacitors are connected in parallel, their equivalent capacitance is given by

$$C_{eq,\ parallel} = \sum_{\text{capacitors}} C_{capacitor} \tag{16.21}$$

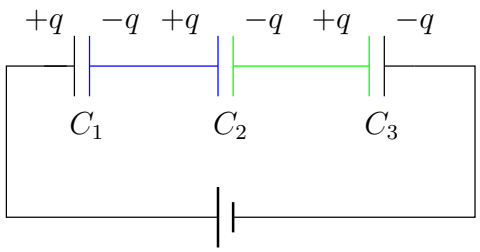


Figure 16.11: Charge on Capacitors in Series

Since no charge crosses the plates of a capacitor, when capacitors are connected in series the interior pairs of plates — those shown in blue and green in fig. (16.11) — are electrically isolated: there is no way for any net charge to jump across to them from adjacent plates. These interior pairs of plates must therefore remain neutral. There is consequently a domino effect: as charge $+q$ flows from the positive terminal of the battery onto the left plate of C_1 in fig. (16.11), the right plate acquires a charge $-q$ by sending $+q$ to the left plate of C_2, and so on — with the result that the plates of all of the capacitors in the series combination end up with the same charges $+q$ and $-q$ across their plates. The voltages across each capacitor will in general differ, but will of course add up to the voltage applied across the series combination:

$$\begin{aligned} V_{applied} &= \sum_{\text{capacitors}} V_{capacitor} \\ &= \sum_{\text{capacitors}} \frac{q}{C_{capacitor}} \\ &= q\left(\sum_{\text{capacitors}} \frac{1}{C_{capacitor}} \right) \end{aligned}$$

If we compare this to $q = CV$ written in the form

$$V_{applied} = \frac{q}{C_{eq,\ parallel}}$$

we see that the parallel combination of capacitors is equivalent to the single capacitance

$$\frac{1}{C_{eq,\ series}} = \sum_{\text{capacitors}} \frac{1}{C_{capacitor}} \tag{16.22}$$

Note that the relations for combining capacitors in series and parallel are the reverse of those for combining resistors in series and parallel. Except for this reversal of rules, circuits involving capacitors in series and parallel are broken down in the same way as circuits involving resistors in series and

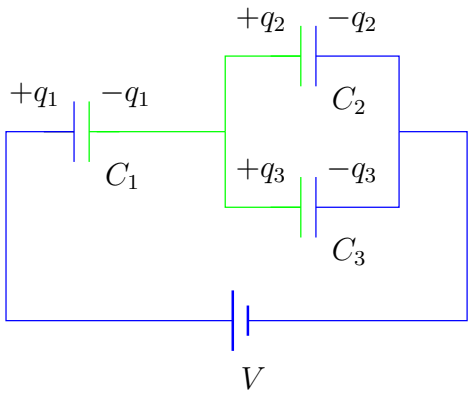


Figure 16.12: Capacitor Loop and Junction Rules

parallel. It is also possible to apply loop and junction rules to capacitor circuits that cannot be resolved into series and parallel combinations: Since the positive plate of the capacitor is at the higher potential, the loop rule will apply with the change in voltage across the capacitor being $+q/C$ if we go from the negative to the positive plate and $-q/C$ if we go from the positive to the negative plate. And the junction rule will take the form that the net charge on any connected set of plates must vanish: since batteries do not supply *net* charge to circuits, the total charge on these plates, which was zero before any batteries were connected, must remain zero after they are connected. Thus for the circuit shown in fig. (16.12), for example, applying the junction rule to the green and blue parts of the circuit yields [18]

$$-q_1 + q_2 + q_3 = 0 \qquad\qquad q_1 - q_2 - q_3 = 0$$

(which are obviously redundant), and from the battery-C_1-C_3 and the C_2-C_3 loops we have

$$-\frac{q_1}{C_1} - \frac{q_3}{C_3} + V = 0 \qquad\qquad -\frac{q_2}{C_2} + \frac{q_3}{C_3} = 0$$

16.5.1 Energy Stored in a Capacitor

Whereas electrical power is dissipated in a resistor, electrical energy is stored in a capacitor. To see how this comes about, imagine starting with an uncharged capacitor C and building up the charge q across its plates in infinitesimal increments dq'. For each such increment, the change in electrostatic potential energy will be

$$dU = dq'\,V'$$

[18] This circuit could, of course, also be broken down into series and parallel combinations, but we're trying to keep the example simple.

where V' is the voltage difference between the plates at that moment. Using $q' = CV'$, we can rewrite this as

$$dU = dq' \, \frac{q'}{C}$$

Since the capacitance C is constant, the only variable on the right-hand side is q', which we can then integrate from zero to the final charge q to obtain the electrical potential energy U stored in the capacitor:

$$U = \int_0^q dq' \, \frac{q'}{C} = \frac{q^2}{2C}$$

Substituting CV for q, we arrive at two equivalent expressions for the electrical energy stored in a capacitor:

$$U = \frac{q^2}{2C} = \tfrac{1}{2}CV^2 \tag{16.23}$$

As we saw in §16.4.1, for a parallel-plate capacitor $C = \epsilon_0 A/d$ and the voltage across the plates is simply $V = Ed$. Since the volume between the plates is Ad, the energy density u (energy per unit volume) between the plates is

$$\begin{aligned} u &= \frac{U}{Ad} \\ &= \frac{\frac{1}{2}CV^2}{Ad} \\ &= \frac{\frac{1}{2}(\epsilon_0 A/d)\,(Ed)^2}{Ad} \\ &= \tfrac{1}{2}\epsilon_0 E^2 \end{aligned} \tag{16.24}$$

As we will see from a more rigorous derivation in §19.3, this actually turns out to be a completely general result for the energy density associated with electric fields.

16.5.2 Games People Play

Certain people — and we all know who *they* are — like to play two kinds of games with parallel-plate capacitors: changing the distance between the plates with the charge held constant, and changing the distance between the plates with the voltage held constant. To win such games, you just have to be very careful about what's constant and what isn't. The charge can be held constant simply by disconnecting the battery before changing the plate separation; the voltage can be held constant by keeping the battery

connected while changing the plate separation, thereby guaranteeing that the voltage across the capacitor always equals the battery's voltage. In either case, changing the distance d between the plates changes the capacitance $C = \epsilon_0 A/d$ — which means, since $q = CV$ must hold throughout, that if q is constant V must change and vice versa.

If, for example, you want to determine the force needed to further separate the plates, you can use eq. (5.13),

$$F = -\frac{dU}{dx}$$

where we are now denoting the separation of the plates by x instead of d in order to avoid the silly-looking dd. In the case where V is held constant, you can take advantage of its constantness by expressing U as $\frac{1}{2}CV^2$, to obtain

$$F = -\frac{d(\frac{1}{2}CV^2)}{dx} = -\frac{\frac{1}{2}(dC)\,V^2}{dx} = -\frac{\frac{1}{2}\,d\Big(\dfrac{\epsilon_0 A}{x}\Big)\,V^2}{dx} = \frac{\epsilon_0 A V^2}{2x^2} \tag{16.25}$$

If instead q is held constant, the more convenient expression for U is $q^2/2C$:

$$F = -\frac{d\Big(\dfrac{q^2}{2C}\Big)}{dx} = -\frac{q^2}{2}\,\frac{d\Big(\dfrac{x}{\epsilon_0 A}\Big)}{dx} = -\frac{q^2}{2\epsilon_0 A} \tag{16.26}$$

Thus an increase in separation (positive dx) yields an outward force (positive F) when V is held constant and an inward force (negative F) when q is held constant. Though the former case may seem weird, the latter case makes sense: we expect not only that the electrical attraction will pull the plates together, but, in the approximation that the plates are essentially infinite, that this force will be constant because the electric field of each plate is constant. More precisely, the field of each plate will be given by eq. (15.7),

$$E = \frac{\sigma}{2\epsilon_0}$$

Since the charge density on each plate is $\sigma = q/A$, the magnitude of the attractive force exerted on the charge q on the other plate will be

$$F = qE = q\,\frac{\sigma}{2\epsilon_0} = q\,\frac{q/A}{2\epsilon_0} = \frac{q^2}{2\epsilon_0 A}$$

which, when you take into account that this force is an attraction, is exactly what we got in (16.26).

To understand how we end up with what seems like a repulsive force between the plates when the voltage is held constant, we have to be mindful

that to keep V constant we need to keep the capacitor connected to a battery. Now, as the separation between the plates increases from x to $x + dx$, the corresponding changes in the capacitance, in the charge across the capacitor, and in the energy stored in the capacitor are[19]

$$dC = d\Big(\frac{\epsilon_0 A}{x}\Big) = -\frac{\epsilon_0 A}{x^2}\,dx$$

$$dq = d(CV) = (dC)\,V = -\frac{\epsilon_0 A}{x^2}\,V\,dx$$

$$dU = d\Big(\tfrac{1}{2}CV^2\Big) = \tfrac{1}{2}\,dC\,V^2 = \tfrac{1}{2}\Big(-\frac{\epsilon_0 A}{x^2}\,dx\Big)\,V^2 = -\frac{\epsilon_0 A V^2}{2x^2}\,dx$$

Thus the capacitance decreases, and in consequence also both the charge on the capacitor and the energy stored in it. But the charge that leaves capacitor has to go somewhere, and the only place it can go is back into the battery. Putting positive charge back into the positive terminal of a battery and negative charge into its negative terminal is in effect charging the battery back up: it involves going uphill in potential energy, to store an additional energy $dU = |dq|\,V$ in the battery. The contribution to our force F from the change in energy of the battery is thus

$$-\frac{dU}{dx} = -\frac{|dq|\,V}{dx} = -\frac{\Big(\dfrac{\epsilon_0 A}{x^2}\,V\,dx\Big)\,V}{dx} = -\frac{\epsilon_0 A V^2}{x^2} \tag{16.27}$$

The total force F between plates is the sum of the contributions (16.25) from the $-dU/dx$ of the capacitor and (16.27) from the $-dU/dx$ of the battery,

$$F_{total} = \frac{\epsilon_0 A V^2}{2x^2} - \frac{\epsilon_0 A V^2}{x^2} = -\frac{\epsilon_0 A V^2}{2x^2}$$

which, being negative, is, as we would have expected, an attractive force.

In summary: our original calculation of the force between plates for the case of constant V was correct, but was only the contribution from the energy stored in the capacitor; there is also a contribution to the force from the energy stored in the battery. Because in this case the energy stored in the capacitor decreases as the plates are moved farther apart, the contribution to the force from the capacitor's energy is repulsive. The increase in the energy stored in the battery more than compensates for this, however, so that overall we end up, as we would expect, doing work against an attractive force to pull the plates apart.

Games similar to those we just played with the separation between the plates can be played with slabs of dielectric being inserted into or pulled

[19] Yes, this last calculation is redundant with what we already worked out above.

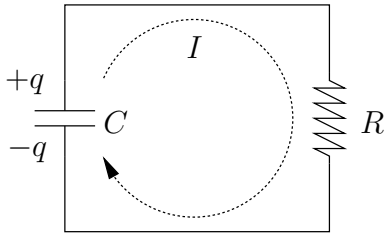


Figure 16.13: A Discharging RC Circuit

out from the region between the plates. These games are also played with $F = -dU/dx$, but with x now representing how far the dielectric is inserted. We will play such games in problems #29 and #30 and look more generally at the relations between the capacitor's and battery's energies in #31.

16.6 RC Circuits

And now at long last we get to a circuit that actually turns out to be useful for something: the DC RC circuit.

First we will look at the case of an initially charged capacitor that is connected across and subsequently discharges through a resistor, as shown in fig. (16.13): positive charge flows from the positive plate of the capacitor clockwise around the circuit, through the resistor, and neutralizes itself against the negative charge on the negative plate of the capacitor. If we apply the loop rule as we go around the circuit with the current, we get a voltage change of $V_c = q_c/C$ as we go from the negative to the positive plate of the capacitor, where V_c and q_c are the voltage and charge across the capacitor. We also get a drop $-IR$ as we cross the resistor with the current. Setting the total change in voltage around the loop to zero, we have

$$\frac{q_c}{C} - IR = 0$$

By definition, the current I is dq/dt. We must, however, be careful to note that the charge dq that passes through the resistor is the charge that *leaves* the capacitor, so that $dq = -dq_c$. In terms of q_c, the current I is therefore $-dq_c/dt$, and our loop relation thus becomes

$$\frac{q_c}{C} + \frac{dq_c}{dt}R = 0$$

We can solve this for q_c as a function of time by separating variables and integrating from the initial charge q_0 across the capacitor plates at time $t = 0$ to the charge q at time t:

$$-\frac{1}{RC}\,dt = \frac{dq_c}{q_c}$$

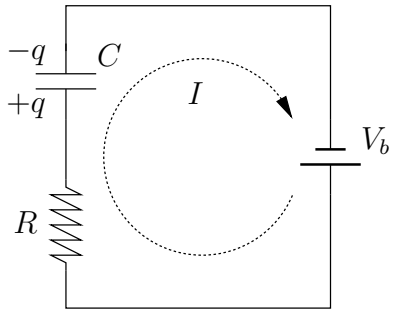


Figure 16.14: A Charging RC Circuit

$$-\int_0^t \frac{1}{RC}\,dt = \int_{q_0}^{q} \frac{dq_c}{q_c}$$

$$-\frac{1}{RC}\,t = \ln q_c\Big|_{q_0}^{q} = \ln\frac{q}{q_0}$$

Taking the exponential of both sides and solving for q, we arrive at

$$q = q_0 e^{-t/RC} \tag{16.28}$$

Solution (16.28) tells us that the charge on the capacitor dies away exponentially, quickly becoming very small but never reaching zero at any finite time. This makes sense physically: the less charge remaining on the capacitor, the smaller the voltage across it and hence across the resistor, with the result that the current flow and rate of discharge of the capacitor decrease.

In each time interval of RC ($1\,\Omega\mathrm{F} = 1\,\mathrm{sec}$), the charge on the capacitor drops by a factor of $1/e$. This time interval is called the time constant of the RC circuit and is usually denoted by τ:

$$\tau = RC$$

Now consider the case of an initially uncharged capacitor connected in series across a resistor and battery of voltage V_b, as shown in fig. (16.14). When the circuit is connected, positive charge flows out of the positive terminal of the battery through the resistor and onto the lower plate of the capacitor. An equal quantity of positive charge leaves the upper plate of the capacitor and flows into the negative terminal of the battery. If we go around the circuit with the current, the loop rule now gives us

$$-IR - \frac{q_c}{C} + V_b = 0 \tag{16.29}$$

This time, the charge dq_c flowing into the capacitor is the same as the charge dq leaving the battery and passing through the resistor. If we use $dq_c = dq$ and $I = dq/dt$, eq. (16.29) becomes

$$-\frac{dq}{dt}\,R - \frac{q}{C} + V_b = 0$$

Although the algebra is a bit more involved this time, again we simply separate variables and integrate: solving for dq/dt yields

$$\frac{dq}{dt} = -\frac{q}{RC} + \frac{V_b}{R} = -\frac{1}{RC}(q - CV_b)$$

and hence

$$\frac{dq}{q - CV_b} = -\frac{1}{RC}\,dt$$

Integrating this as the charge on the capacitor goes from zero at time $t = 0$ to q at time t, we have

$$\begin{aligned}
\int_0^q \frac{dq}{q - CV_b} &= -\int_0^t \frac{1}{RC}\,dt \\
\ln(q - CV_b)\Big|_0^q &= -\frac{1}{RC}\,t \\
\ln\Big(\frac{q - CV_b}{-CV_b}\Big) &= -\frac{1}{RC}\,t
\end{aligned}$$

which, when solved for q, gives

$$q = CV_b\big(1 - e^{-t/RC}\big) \tag{16.30}$$

Our solution (16.30) tells us that the charge across the capacitor builds up exponentially toward the asymptotic value CV_b, with the rate of buildup slowing as this full charge is approached. This again makes sense physically: as charge and hence voltage build up across the capacitor, more of the applied voltage V_b of the battery is across the capacitor and less across the resistor, with the result that the current flow and rate at which charge is building up on the capacitor decrease. In the limit $t \to \infty$, all of the applied voltage is across the capacitor and none across the resistor, so that no current flows through the resistor and the charge on the capacitor is $q = CV_b$ — just as it would be if the capacitor were hooked up directly across the battery without any resistor in the circuit. The time constant for the buildup is the same $\tau = RC$ that it was for the discharging RC circuit.

Believe it or not, this charging RC circuit actually has a practical application: it can be used to make lights blink, like those flashers you see on highway warning barrels. All you need to do is connect a neon light-bulb in parallel with the capacitor, as shown in fig. (16.15). When the charge and voltage across the capacitor are small, the bulb has effectively infinite resistance, so that no current flows through it; it is unlit and effectively not even a part of the circuit. But when the charge and voltage across the capacitor reach a certain threshold value, the effective resistance of the bulb suddenly drops to zero and the charge that has accumulated across the capacitor zips

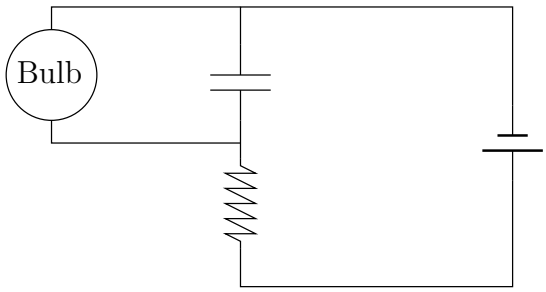


Figure 16.15: The Old Blinking-Barrel Trick

through the bulb in the form of a spark that gives a flash of light. This discharges the capacitor, returning the circuit to its initial state, so that the charging process starts over again. By adjusting the value of the time constant RC, the interval between flashes can be made as long or short as desired.

Or at least that's the way these flashing lights used to work. These days, everything is digital.

16.7 Treating AC As DC

We have now managed to get almost all the way through the chapter without ever addressing the term DC that constitutes fully half of its title. If that has been causing you angst, you are about to obtain closure on this issue: *DC* stands for direct current, the current that flows through circuits connected to sources of constant voltage such as batteries.

Household electricity is, however, AC, the alternating current that flows in response to applied voltages that vary sinusoidally. Although we will not deal with AC circuits in detail until Chapter 20, it turns out, as you might expect, that in AC circuits the voltages, currents, and charges all vary sinusoidally at the same frequency as the applied voltage. For example, the current flowing through a circuit in response to an applied voltage of angular frequency ω is of the form

$$I = I_{max} \sin(\omega t + \varphi) \tag{16.31}$$

where I_{max} is the amplitude (that is, the peak value) of the current and φ is its phase relative to the applied voltage. This phase takes into account that in general the current does not peak at the same time as the applied voltage.

These oscillations of currents, voltages, and charges at differing phases make calculations for AC circuits considerably more involved and messy than those for DC circuits. But our objective at this point is very limited: we would merely like to find some way of applying Ohm's bogus law and the power relation $P = IV = I^2R = V^2/R$ to AC circuits.

Your first thought might be to obviate the complications of time variation and phases by working with average values. Unfortunately, this approach will fail because the average value of any quantity that oscillates sinusoidally of course vanishes: the time average of the current of eq. (16.31), for example, over each full cycle of period $T = 2\pi/\omega$ is

$$\begin{aligned} I_{avg} &= \frac{1}{T}\int_0^T I_{max}\sin(\omega t+\varphi)\,dt \\ &= \frac{I_{max}}{\omega T}\int_0^{\omega T}\sin(\omega t+\varphi)\,d(\omega t) \\ &= \frac{I_{max}}{2\pi}\int_0^{2\pi}\sin(\theta+\varphi)\,d\theta \\ &= 0 \end{aligned}$$

Physically, this comes about because the currents, voltages, etc., reverse themselves in the second half of each cycle (and hence the term *alternating current*).

Your next thought might be to work in terms of the peak values of the currents and voltages. And it does in fact turn out that the currents and voltages are in phase with each other in circuits involving only resistors, so that Ohm's bogus law will hold true for their peak values in the form $V_{max} = I_{max}R$, and likewise $P_{max} = I_{max}V_{max}$. But we are usually concerned with the energy consumed by a circuit over a period of time, and this is most directly related, not to the peak, but to the average power. When using peak values, we would have to be mindful that the average power consumed by the circuit is half of P_{max}: since $P = IV$ varies as $\sin^2\omega t$,

$$\begin{aligned} P_{avg} &= \frac{1}{T}\int_0^T P_{max}\sin^2\omega t\,dt \\ &= \frac{P_{max}}{2\pi}\int_0^{2\pi}\sin^2\theta\,d\theta \\ &= \tfrac{1}{2}P_{max} \\ &= \tfrac{1}{2}I_{max}V_{max} \end{aligned} \tag{16.32}$$

There is, however, a better way: if we work in terms, not of peak, but of root-mean-square (rms) values, then we don't even have to remember the factor of $\frac{1}{2}$ in the average power. As the name indicates, the root-mean-square value of a quantity is the square root of the mean of its square. For a quantity that varies sinusoidally, the root-mean-square value will be the peak value divided by $\sqrt{2}$. For example, for a current I of eq. (16.31) the

root-mean-square value I_{rms} would be [20]

$$\begin{aligned} I_{rms} &= \sqrt{\frac{1}{T}\int_0^T I_{max}^2 \sin^2(\omega t + \varphi)\, dt} \\ &= I_{max}\sqrt{\frac{1}{T}\int_0^T \sin^2(\omega t + \varphi)\, dt} \\ &= I_{max}\sqrt{\tfrac{1}{2}} \end{aligned}$$

As it will for peak values, Ohm's bogus law will hold true for the root-mean-square values of the current and voltage in the form

$$V_{rms} = I_{rms} R$$

And since

$$I_{rms} V_{rms} = \frac{I_{max}}{\sqrt{2}} \frac{V_{max}}{\sqrt{2}} = \tfrac{1}{2} I_{max} V_{max}$$

we also conveniently have, by eq. (16.32),

$$P_{avg} = I_{rms} V_{rms}$$

And so if we work with the root-mean-square values of the voltages and currents, not only can we apply Ohm's bogus law to AC circuits, but we can also obtain the average power by the usual $P = IV$ (and its variants I^2R and V^2/R). In fact, in the context of household electricity it is commonly the root-mean-square values of voltages and currents that are quoted: when an outlet is said to be 120 V, that means that its root-mean-square voltage is 120 V; the peak voltage would be $120\sqrt{2} \approx 170$ V.

[20] That the average value of $\sin^2$ or $\cos^2$ over any whole number of sinusoidal "humps" is $\frac{1}{2}$ is one of those very handy things to remember. Not likely to impress a potential date at a cocktail party, but it does crop up in calculations more frequently than you might expect.

16.8 Problems

1. A current I flows radially outward from a point evenly in all directions. Determine the current density $\mathbf{j}$ at a distance r from the point.

2. For the current density $\mathbf{j} = j_0\,\hat{\mathbf{z}}$, where j_0 is a positive constant, show explicitly that you get the same result for the current through

   (a) A circle of radius a, centered at the origin, that lies in the xy plane.

   (b) The hemisphere with the same perimeter as that circle.

3. A wire of radius a carries a current that varies over its cross section as

$$j = j_0\left(\frac{r}{a}\right)^n$$

   where j_0 is a constant.

   (a) Determine the value of j_0 in terms of the current I_0 flowing down the wire. What restrictions are there on the value of n?

   (b) Explain how the dependence of your result for j_0 on n makes sense physically.

   (c) Explain the physical pathology of those cases (that is, those values of n) that do not yield a solution.

4. The belt of a Van de Graf generator[21] has width a, moves at speed v, and carries a surface charge density σ. Determine the rate at which the belt is conveying charge to the dome.

5. A hair drier designed for a 120 V outlet is rated at 1500 W.[22]

   (a) How much current does the hair drier draw from the outlet?

   (b) What is the resistance in the coils in the hair drier?

---

[21] If you have no idea what a Van de Graf generator is, see footnote 26 on p.817. You still won't have much idea what a Van de Graf is, but it's better than nothing.

[22] All electrical appliances should have a label indicating the voltage for which they are designed and the current and power they draw; if you're curious about an appliance, you just have to find this label. Next to the fuses in the fuse boxes (or circuit breakers in the breaker boxes) you will find the maximal number of amps each line can safely carry. And, unless it has been torn or worn off, every extension cord has a tag indicating how much current it can safely carry.

For appliances such as air conditioners that draw variable currents, you can buy power (watt) meters that will measure instantaneous current and power and total energy consumed over a period of time. Usually these meters, like your electric bill, report this consumed energy in kilowatt-hours (kWh): $1\,\text{kWh} = 1000\,\text{W} \times 1\,\text{hr} = 3.6 \times 10^6\,\text{J}$.

On treating AC household voltage as DC, see §16.7.

6. A battery charger for a 9 V battery supplies 140 mA.

   (a) How much power does the charger supply to the battery?

   (b) How much charge does the charger put through the battery each hour?

   (c) How do you reconcile this result with the fact that 1 C is a huge quantity of charge — about what you'd get in a typical backyard lightning bolt? Household appliances sometimes draw even 10 or 20 A, which is 10 or 20 C *per second*. So why isn't it like a scene from a Frankenstein movie every time you use your toaster, with huge blue lightning bolts spraying out all over the kitchen? [23]

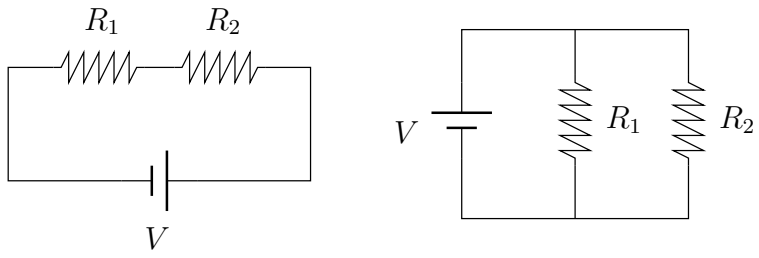


Figure 16.16: Problem 7

7. (a) For the circuits shown on each side of fig. (16.16), determine

      i. The voltage across, current through, and power dissipated in each resistor.

      ii. The current and power supplied by the battery.

   (b) In what ways are your results for the current and power supplied by the battery consistent with your results for the voltages across, currents through, and power dissipated in the two resistors?

8. (a) How does the value of the equivalent resistance compare to the values of the individual resistances when resistors are connected

      i. In series.

      ii. In parallel.

   (b) Make physical sense of this.

[23] How cool would that be? But unfortunately life just isn't that exciting. As all of you know who've ever asked for a flame-thrower for your birthday or tried to get permission to toss the whole can of lighter fluid into the barbecue and make a giant fireball. Ah, well.

9. Having neglected to read the safety instructions, Curly [24] tries to use a curling iron while standing, soaking wet, in a tub full of water. Water dripping into the curling iron causes a short circuit, with the result that the voltage V_0 of the wall outlet is connected *in parallel* across Curly (resistance R), his damp towel (resistance $\frac{2}{3}R$), and his "rubber" ducky (resistance also $\frac{2}{3}R$, because, like all children's toys, it is made out of materials that are highly flammable, give off toxic fumes when burned, and conduct electricity well enough to cause fires [25]).

   (a) Draw a diagram of the circuit.

   (b) Determine the voltage across, current through, and power dissipated in each of the three resistances (Curly, the towel, and the rubber ducky).

   (c) Show, in terms of the power drawn from the wall outlet and the power dissipated in each of the three resistances, that energy is conserved in this circuit.

10. (Inspired by *National Lampoon's Christmas Vacation*.) While two of her kittens (each resistance R) are nursing from her, a mother cat (resistance $2R$) chews through the extension cord that runs from the wall outlet (voltage V_0) to the TV. Together, they form an electrical circuit in which the two kittens are connected in parallel and this parallel combination of the kittens is in series with the mother.

    (a) Draw a diagram of the circuit.

    (b) Determine the voltage across, current through, and power dissipated in each of the three resistances (the two kittens and the mother).

    (c) Show, in terms of the power drawn from the wall outlet and the power dissipated in each of the three resistances, that energy is conserved in this circuit.

11. For each of the combinations of resistors in fig. (16.17):

    (a) Determine the equivalent resistance of the combination.

    (b) Determine the current and power delivered by the battery.

    (c) Determine the voltage across, current through, and power dissipated in each resistor.

[24] Referring, of course, to the Curly of *The Three Stooges* fame — we needed some bald person certain to be familiar to all educated and cultured persons.

[25] As we know from *The Simpsons*, children's toys make the best fuel for a bonfire.

Which reminds us of another gem from *The Simpsons*, when Nelson, in shocked disillusionment upon learning, along with the other boys, that Krusty the Klown is a felon, declares "Oh, man, I don't believe in anything anymore ... I'm going to law school."

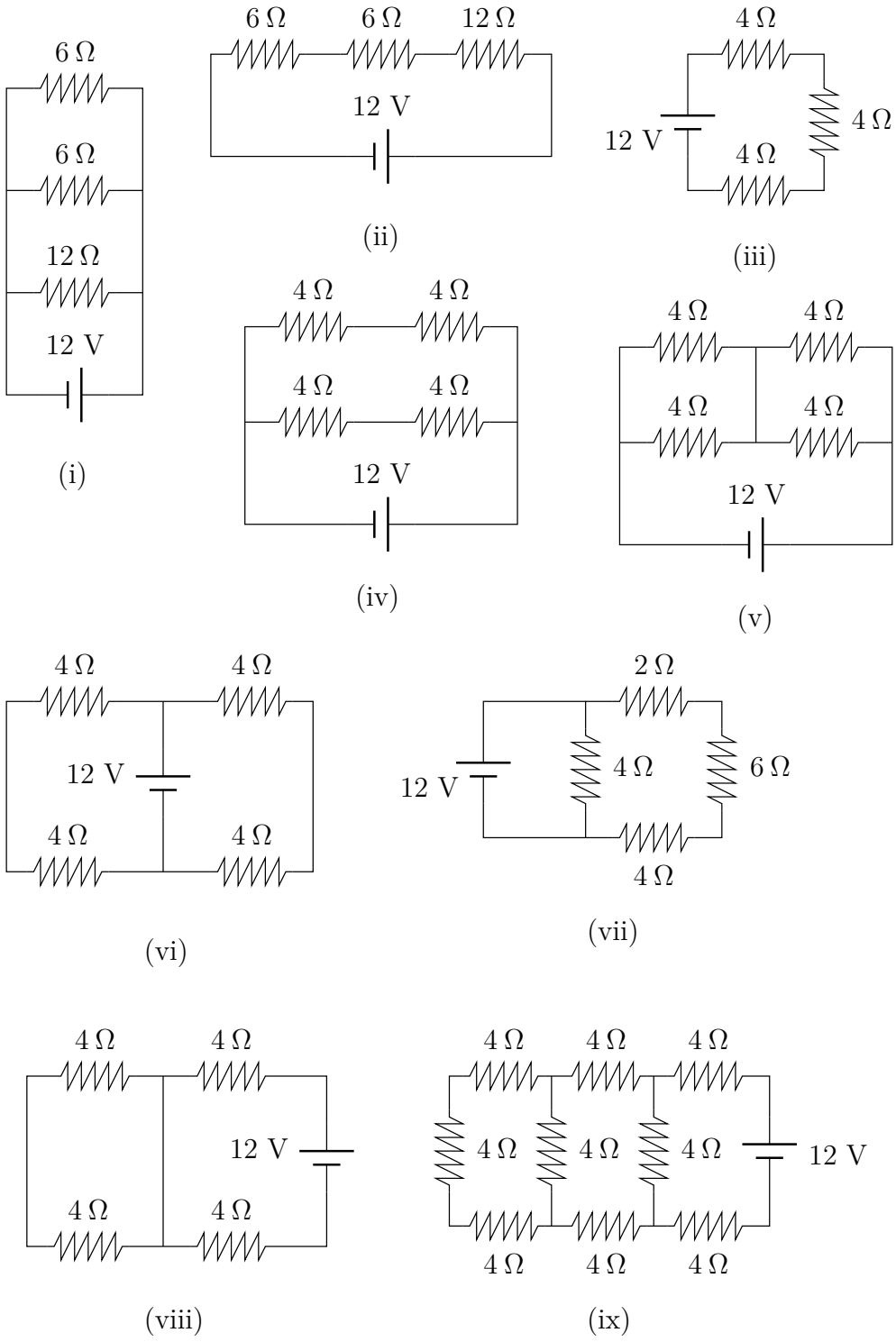


Figure 16.17: Problem 11

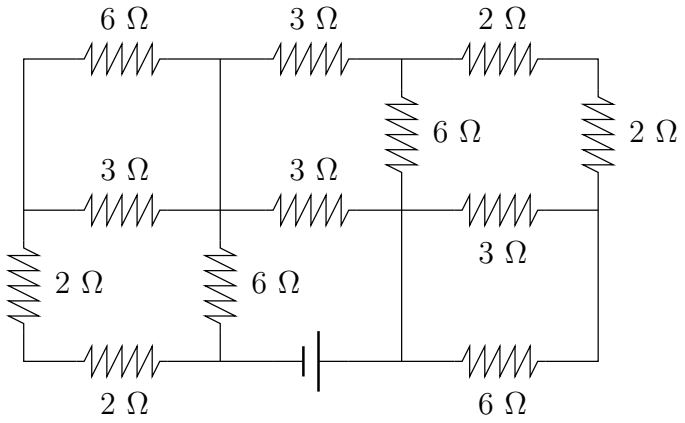


Figure 16.18: Problem 12

12. Determine the equivalent resistance of the circuit shown in fig. (16.18).

    (If done right, this doesn't require all that much work.)

13. In addition to their resistance, resistors are normally rated by the maximal power dissipation they can sustain; if the power dissipated exceeds this upper limit, the heat generated in the resistor will cause it to burn up.[26]

    Suppose you have a limitless supply of $2\,\Omega$, $\frac{1}{4}\,\mathrm{W}$ resistors.

    (a) Can't you just feel everyone's envy?

    (b) How can you combine these to construct the equivalent of a $2\,\Omega$, $1\,\mathrm{W}$ (or better) resistor?

    (c) Is your answer to the preceding part unique? Are there a finite number of other solutions? An infinite number?

    (d) How can you combine these to construct the equivalent of a $\frac{1}{5}\,\Omega$ resistor?

    (e) How can you combine these to construct the equivalent of a $\frac{1}{5}\,\Omega$ resistor that can withstand $5\,\mathrm{W}$ or better?

14. A resistance R is connected across the terminals of a battery of unknown voltage. Determine the relationship between R and a second resistance R' if

    (a) The current drawn from the battery halves when R' is connected in series with R.

    (b) The current drawn from the battery triples when R' is connected in parallel with R.

[26] While we're at it, capacitors are likewise rated by a maximal voltage (the break-down voltage) they can sustain before a spark can leap between plates, not only discharging the capacitor but ruining it.

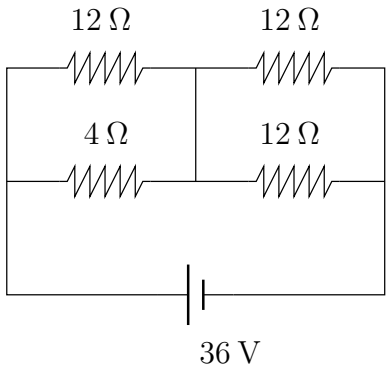


Figure 16.19: Problem 15

15. Determine the current in each branch of the circuit shown in fig. (16.19), including the "cross piece."

16. You have four resistors, each of resistance R. Figure out all the electrically distinct ways that these four resistors can be connected. Each combination must use all four resistors. For each combination, draw the circuit diagram and determine the equivalent resistance. See the footnote if you need a hint.[27]

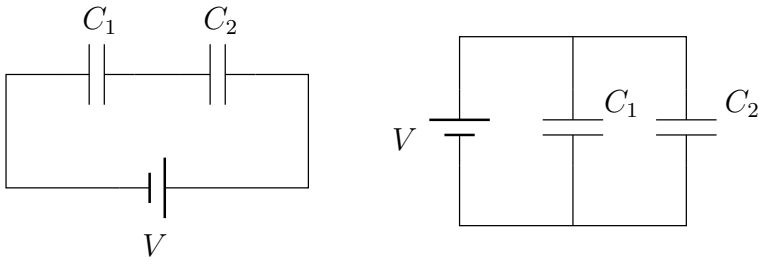


Figure 16.20: Problem 17

17. (a) For the circuits shown on each side of fig. (16.20), determine

    i. The voltage and charge across, and the energy stored in, each capacitor.

    ii. The charge and energy supplied by the battery.

    (b) In what ways are your results for the charge and energy supplied by the battery consistent with your results for the charges across and the energy stored in the two capacitors?

[27] There are ten distinct combinations, two of which just happen by coincidence to have the same value for the equivalent resistance. Think about how you can construct nested series and parallel combinations.

18. (a) How does the value of the equivalent capacitance compare to the values of the individual capacitances when capacitors are connected

    i. In series.

    ii. In parallel.

    (b) Make physical sense of this.

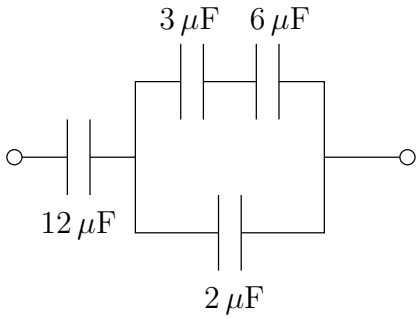


Figure 16.21: Problem 19

19. Just in case you encounter such animals in your travels, the two little loopy bits in fig. (16.21) are the symbol some books use for the ends of the wires by which you would hook this combination of capacitors up to the terminals of a battery. Anyway, determine the charge and voltage across each capacitor if this combination is connected across a 12 V battery.

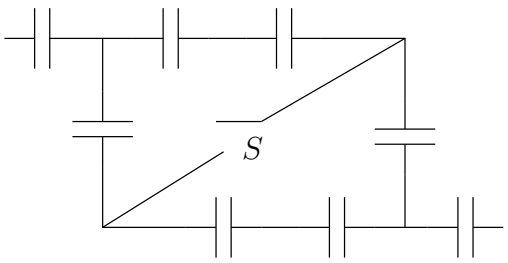


Figure 16.22: Problem 20

20. In fig. (16.22), all of the capacitors have the same value C, and S represents an open switch.

    (a) Determine the equivalent capacitance of this utterly useless arrangement of capacitors.

    (b) If the switch S were now closed, how would this change your result for the equivalent capacitance?

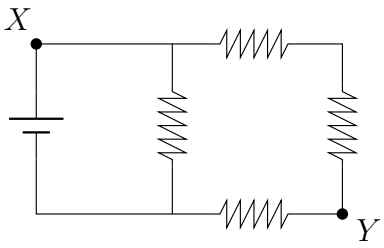


Figure 16.23: Problem 21

21. In the circuit shown in fig. (16.23), the resistors are each $4\,\Omega$ and the battery is $12\,\mathrm{V}$.

    (a) Determine the voltage difference between the points X and Y.

    (b) Which point is at the higher voltage?

    (c) How would your result change if the resistors were all $3\,\Omega$ rather than $4\,\Omega$?

    (d) Explain why you can or cannot say that the voltage at point X is $12\,\mathrm{V}$.

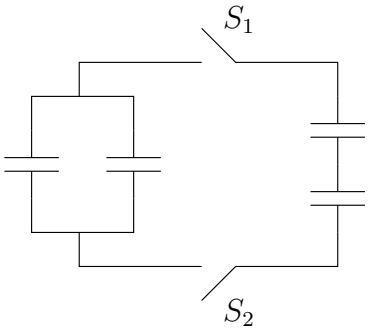


Figure 16.24: Problem 22

22. In the circuit shown in fig. (16.24), all the capacitors have the same capacitance C. The parallel combination on the left side is initially charged to voltage V_0. The series combination on the right is initially uncharged.

    (a) The switch S_1 is closed, but S_2 is left open. Explain why nothing happens.

    (b) Now both switches are closed. Describe what happens and determine the final charges on each of the four capacitors.

    (c) Suppose that instead the parallel pair on the left side were initially charged to voltage V_0 and the series pair on the right were initially charged to $-V_0$ (that is, to V_0, but facing the opposite way, so that if the upper plates of the parallel pair were positive, the lower plates of each of the series pair would be positive). Determine the final charges on each of the four capacitors.

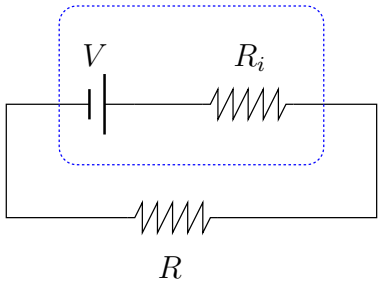


Figure 16.25: Problem 23

23. To a very rough approximation, an ordinary battery behaves as though it were not just a pure voltage, but a voltage in series with an internal resistance, as shown in fig. (16.25): R represents the external load across which the battery is connected, and the battery itself consists, as indicated by the dotted blue box, of its pure voltage V in series with its internal resistance R_i. When you are drawing current from the battery, the voltage across its terminals — that is, the voltage across the blue box in fig. (16.25) — is therefore less than the value printed on the label because some of the battery's pure voltage is dropped across the internal resistance. As the battery wears out, this internal resistance gradually increases.

    Suppose an old battery with a labeled voltage V actually delivers only voltage V_1 while supplying current I_1 to a flashlight bulb.

    (a) Determine the internal resistance of the battery.

    (b) Determine the voltage across the battery's terminals if you now draw a current I_2 from it.

    (c) What is the maximal current that the battery can supply?

    (d) Determine the relation for the current I delivered to an external load R by a battery with labeled voltage V and internal resistance R_i. Show that the power delivered to the load is

$$P_{load} = \left(\frac{V}{R + R_i}\right)^2 R$$

    (e) Show that maximal power is delivered to the load when $R = R_i$.

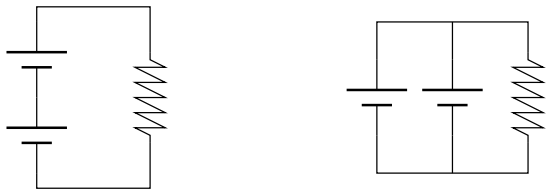


Figure 16.26: Problem 24

24. Batteries that drive circuits are sometimes connected in series or parallel, as shown in fig. (16.26).

    (a) What effect would connecting two batteries in series have? What purpose would this serve?

    (b) What effect would connecting two batteries in parallel have? What purpose would this serve?

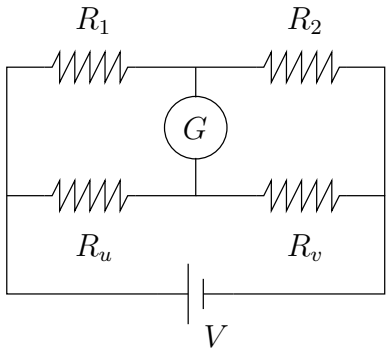


Figure 16.27: Problem 25

25. An ammeter, as the name would suggest, measures current, and a galvanometer is a very sensitive ammeter: if there is even a tiny voltage difference across the galvanometer, the resulting current through it will cause its needle to deflect visibly. Fig. (16.27) shows a device known as a Wheatstone bridge that can be used to determine the value of an unknown resistance: the circled G symbolizes a galvanometer, R_1 and R_2 are known resistances, and R_u is the unknown resistance. R_v represents a variable resistor, which, as you might have guessed, is a resistor with an adjustable resistance.[28]

    Show that when the value of a variable resistance has been adjusted so that there is no voltage difference across the galvanometer, the unknown resistance is given by

$$R_u = \frac{R_1}{R_2} R_v$$

    See the footnote if you need a hint.[29]

[28] Constructing a variable resistor is actually pretty simple. Remind me to haul one out.

[29] Think about the currents flowing through the various resistors as well as the volt-

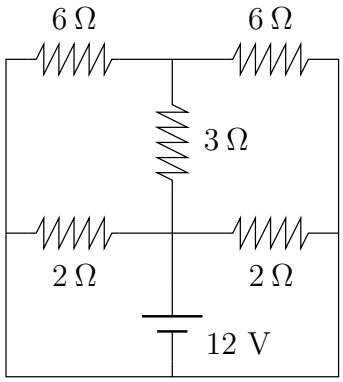


Figure 16.28: Problem 26

26. (a) Analyze the circuit shown in fig. (16.28) by loop and junction rules. Naively, there are eight branches and therefore eight currents to contend with. But if you make good use of the symmetry of the circuit, you can immediately reduce this to five currents and in fact won't have to deal with more than two simultaneous equations in two unknowns.

    (b) Show that the resistors are therefore equivalent to a single resistance of $\frac{6}{7}\ \Omega$.

27. (a) A cube is constructed out of coat hangers. Each edge has a resistance R. Determine the equivalent resistance between corners at opposite ends of a long diagonal. See the footnote if you need a hint.[30]

    (b) Do the same for a four-dimensional cube.

    (c) Do the same for a ten-dimensional cube.

28. (a) Determine the force that the plates of a capacitor exert on each other in each of the following cases:

        i. A parallel-plate capacitor with plates of area A that are a distance ℓ apart and carry charge $\pm q$.

        ii. A cylindrical capacitor with plates of radii a and b that carry charge $\pm q$.

        iii. A spherical capacitor with plates of radii a and b that carry charge $\pm q$.

    (b) What, if anything, do your answers tell you about how much force you would need to overcome, in the sense of work per unit length, to increase the separation between the plates in each of the preceding cases?

---

ages across them. That there is no voltage difference across the galvanometer tells you something about these currents.

[30] Think about how the current entering one corner of the cube divides as it travels the various paths to the far corner.

Figure 16.29: Problem 29

29. A parallel-plate capacitor has plates of area A separated by a distance ℓ. A slab of dielectric of dielectric constant κ with the same area A and thickness ℓ is placed between the plates, as shown from a side view in fig. (16.29). Determine

    i. The change in the energy stored in the capacitor,

    ii. The work that you do, and

    iii. The change in the energy stored in the battery (for #29b only)

    when this slab is inserted while

    (a) Keeping the charge $\pm q$ on each plate constant.

    (b) Keeping the voltage V between the plates constant.

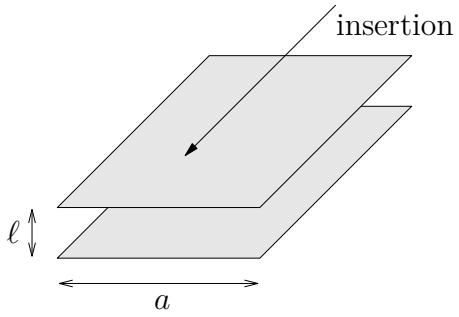


Figure 16.30: Problem 30

30. Fig. (16.30) shows the parallel-plate capacitor of #29 from an oblique angle. The separation between the plates and area of the plates are again ℓ and A, respectively, and now the width of the plates is a.

(a) Show that when the charge on the plates is held constant the force exerted on the dielectric slab as it enters the region between the plates is

$$F = \frac{q^2}{2\epsilon_0 A^2}\, a\ell\, \frac{\kappa - 1}{\kappa}$$

You may find it helpful to note that in terms of the additional distance dx the slab is inserted and the energy U stored in the capacitor, the force is

$$F = -\frac{dU}{dx} = -\frac{dU}{a\ell\, dx}\, a\ell = -\Delta u\, a\ell$$

where Δu is the change in the energy density $\frac{1}{2}\epsilon_0 E^2$ as the dielectric moves forward a distance dx and the volume $a\ell\, dx$ goes from being empty to being filled with dielectric. See the footnote if you need a further hint.[31]

(b) Show that when the voltage is held constant by means of a battery of voltage V, the same force is exerted on the dielectric slab. This time, you may find it helpful to note that the battery's contribution to the force is

$$-\frac{dU}{dx} = -\frac{dq\, V}{dx}$$

where dq is the charge that flows from the battery onto the plates as the dielectric moves forward a distance dx. This charge dq may in turn be expressed as $\Delta\sigma\, dA$, where dA is the additional area of the plates with dielectric in between and $\Delta\sigma$ is the change in the charge density due to the presence of the dielectric. See the footnote if you need a further hint.[32]

[31] As usual, $\epsilon_0 \to \kappa\epsilon_0$ in the presence if the dielectric. What therefore happens to the field $E = \sigma/\epsilon_0$ when the charge is held constant?

[32] When evaluating the capacitor's contribution $-\Delta u\, a\ell$ to the force, be mindful that the constant voltage V between the plates now entails, by $V = |-\int \mathbf{E}\cdot d\mathbf{r}|$, constant E. Since $E = \sigma/\epsilon_0$ is constant, how does σ change?

31. Show that, quite generally, when a change dC is made in the capacitance of a capacitor, the changes dU_q and dU_V in the energy stored in the capacitor when the charge q and voltage V are held constant, respectively, and in the latter case the change dU_b in the energy stored in the battery, are related by

$$dU_V = -dU_q \qquad\qquad dU_b = 2\,dU_q$$

so that we always have

$$dU_V + dU_b = 2\,dU_q - dU_q = dU_q$$

See the footnote if you need a hint.[33]

32. Suppose you want to construct a blinking-light circuit like that of fig. (16.15) on p.868 and have at your disposal a 12 V battery, a bulb that flashes when the voltage across it reaches 10 V, a 1 μF capacitor, and a variety of resistors. What resistance should you use in order to make the light flash every 3.0 sec?

33. (a) For a discharging RC circuit, show explicitly that
    i. At any given instant the power dissipated in the resistor equals the rate at which the capacitor is losing energy.
    ii. The energy initially stored in the capacitor equals the total energy dissipated in the resistor as the capacitor fully discharges.

    (b) For a charging RC circuit, show explicitly that
    i. At any given instant power is conserved.
    ii. The total energy dissipated in the resistor and the energy ultimately stored in the capacitor account for the total energy supplied by the battery as the capacitor goes from being uncharged to being fully charged.

[33] Consider the changes in

$$U = \frac{q^2}{2C} \qquad\qquad U = \tfrac{1}{2}CV^2 \qquad\qquad U = qV$$

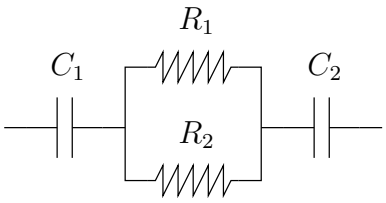


Figure 16.31: Problem 34

34. An RC circuit is constructed by connecting a battery across the combination of resistors and capacitors shown in fig. (16.31). Determine the time constant of this RC circuit.

16.9 Sketchy Answers

(1) $\mathbf{j} = \dfrac{I}{4\pi r^2}\,\widehat{\mathbf{r}}$.

(2) In both cases, you should get $j_0 \pi a^2$.

(3a) $j_0 = \dfrac{I_0}{2\pi a^2}(n+2)$.

(4) $I = \sigma a v$.

(5a) 12.5 A.

(5b) 9.6 Ω.

(6b) 504 C.

(6a) 1.26 W.

(11)

Part	R_{eq} (Ω)	Set of Voltages (V)
i	$\frac{12}{5}$	*Duh!*
ii	24	3, 6
iii	12	4
iv	4	6
v	4	6
vi	4	6
vii	3	2, 4, 6, 12
viii	8	0, 6
ix	$\frac{164}{15}$	You'd have to be nuts.

(12) 5 Ω.

(13) You're on your own on this one. Although if your answer to the first part was yes, you might consider getting some sort of help.

(14a) $R' = R$.

(14b) $R' = \frac{1}{2}R$.

(15) The set of currents includes 1, 2, 3, and 4 A.

(19) The sets of charges and voltages are 18 and 36 μC and 3, 6, and 9 V.

(20a) $\frac{2}{7}C$.

(20b) $\frac{1}{5}C$.

(21a) 8 V.

(22b) The set of charges includes $\frac{2}{5}CV_0$ and $\frac{4}{5}CV_0$.

(22c) The set of charges includes $\frac{3}{10}CV_0$ and $\frac{3}{5}CV_0$.

(23a) $\dfrac{V - V_1}{I_1}$.

(23b) $V - (V - V_1)\dfrac{I_2}{I_1}$.

(23c) $\dfrac{V}{V - V_1}\, I_1$.

(26a) The set of currents is 1, 2, 6, 7, and 14 A.

(27a) $\frac{5}{6}R$.

(27b) $\frac{2}{3}R$.

(27c) $\frac{487}{2160}R$.

(28(a)i) $\dfrac{q^2}{2\epsilon_0 A}$.

(29a) The change in the energy stored in the capacitor is $-\dfrac{q^2\ell}{2\epsilon_0 A}\dfrac{\kappa - 1}{\kappa}$.

(29b) The change in the energy stored in the capacitor is $\dfrac{\epsilon_0 A V^2}{2\ell}(\kappa - 1)$.

The energy supplied by the battery is $\dfrac{\epsilon_0 A V^2}{\ell}(\kappa - 1)$.

(32) $\dfrac{3}{\ln 6}\,\mathrm{M}\Omega$.

Chapter 17

Magnetostatics

In electrostatics we were concerned with the electric fields of static distributions of charge. Because no charge was in motion, no magnetic fields were generated, nor was there any time dependence to the electric field. Thus in electrostatics the Maxwell equations

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\,\rho \qquad\qquad \nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

$$\nabla \cdot \mathbf{B} = 0 \qquad\qquad \nabla \times \mathbf{B} = \mu_0\,\mathbf{j} + \epsilon_0\mu_0\frac{\partial \mathbf{E}}{\partial t}$$

reduced to

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\,\rho \qquad\qquad \nabla \times \mathbf{E} = 0 \tag{17.1}$$

The former of eqq. (17.1), Gauss's law, told us that electric charge is the source of electric field. In applications, its integral form

$$\oint_S dA\,\hat{\mathbf{n}} \cdot \mathbf{E} = \frac{1}{\epsilon_0}\,q_{enclosed}$$

enabled us to solve for the electric fields of highly symmetric distributions of charge. The latter of eqq. (17.1) told us that the electrostatic field is a pure gradient, that is, is of the form $\mathbf{E} = -\nabla\varphi$, which led us to the electrostatic potential

$$\varphi = -\int \mathbf{E} \cdot d\mathbf{r}$$

and to the electrostatic potential energy

$$U = q\varphi$$

The relation $\mathbf{E} = -\nabla\varphi$ also provided us with an additional method of solving for the electric field.

In magnetostatics we are concerned with time-independent distributions of current and time-independent magnetic fields. Together, these two restrictions dictate that magnetostatic currents not only not vary with time, but also form closed loops that do not redistribute charge:[1] Any current that were to redistribute charge would give rise to a charge density ρ that varied with time. By Gauss's law

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\,\rho$$

this would lead to an electric field that would vary with time, and by Ampère's law

$$\nabla \times \mathbf{B} = \mu_0\,\mathbf{j} + \epsilon_0\mu_0\frac{\partial \mathbf{E}}{\partial t}$$

there would in turn be a nonstatic contribution to the magnetic field from the $\partial\mathbf{E}/\partial t$ term.

The two Maxwell equations involving the magnetic field, namely the magnetic Gauss's law

$$\nabla \cdot \mathbf{B} = 0$$

and Ampère's law

$$\nabla \times \mathbf{B} = \mu_0\,\mathbf{j} + \epsilon_0\mu_0\frac{\partial \mathbf{E}}{\partial t} \tag{17.2}$$

reduce in magnetostatics to

$$\nabla \cdot \mathbf{B} = 0 \qquad\qquad \nabla \times \mathbf{B} = \mu_0\,\mathbf{j} \tag{17.3}$$

In §2.7.1 we saw that the vanishing of the divergence of the magnetic field $\mathbf{B}$ implies that $\mathbf{B}$ is a pure curl, that is, that $\mathbf{B}$ can be expressed as $\mathbf{B} = \nabla \times \mathbf{A}$. The latter of eqq. (17.3), Ampère's law proper,[2] tells us that electric current gives rise to magnetic field. Its equivalent integral form

$$\oint_C \mathbf{B} \cdot d\mathbf{r} = \mu_0\, I_{enclosed} \tag{17.4}$$

will enable us to determine the magnetic field of certain simple distributions of current. We will also find that, analogous to using $\mathbf{E} = -\nabla\varphi$ to solve for the electric field, there is a method of solving for the magnetic field by means of $\mathbf{B} = \nabla \times \mathbf{A}$.

But before we get into techniques for calculating magnetic fields, a few words about magnetic forces are in order.

[1] Constant currents that start and end out at infinity are, however, allowed, because one can make such a current into a closed loop by an arc of current out at infinity. Since even a magnetic monopole field would, as it turns out, fall off as $1/r^2$, the addition of this arc would not affect the value of the magnetic field at finite locations. Thus an infinite straight line of current effectively constitutes a closed loop. A semi-infinite line of current, however, would not.

[2] Recall that, strictly speaking, Ampère's law is $\nabla \times \mathbf{B} = \mu_0\,\mathbf{j}$, without the $\partial\mathbf{E}/\partial t$ term. We have been a bit loose in referring to eq. (17.2) as Ampère's law.

17.1 Magnetic Forces

Recall that the magnetic force on a charge q moving with velocity $\mathbf{v}$ is given by eq. (14.14),

$$\mathbf{F} = q\mathbf{v} \times \mathbf{B}$$

where $\mathbf{B}$ is the magnetic field at the location of the charge q. The magnitude of this magnetic force is

$$F = qvB \sin\theta \tag{17.5}$$

where θ is the angle between the charge's velocity $\mathbf{v}$ and the magnetic field $\mathbf{B}$. This magnetic force law has several significant properties:

- The direction of the magnetic force is always perpendicular to both the charge's velocity and the magnetic field.
- To the extent that the charge's velocity is along the magnetic field — that is, parallel or antiparallel to it — , there is no magnetic force; you only get a magnetic force from the component of the velocity perpendicular to the magnetic field.
- Since the magnetic force is always perpendicular to the charge's velocity $\mathbf{v} = d\mathbf{r}/dt$ and hence to its displacement $d\mathbf{r}$, magnetic forces never do any work:

  $$dW = \mathbf{F} \cdot d\mathbf{r} = q(\mathbf{v} \times \mathbf{B}) \cdot d\mathbf{r} = q\left(\frac{d\mathbf{r}}{dt} \times \mathbf{B}\right) \cdot d\mathbf{r} = 0$$

  because $d\mathbf{r} \times \mathbf{B}$ is always perpendicular to $d\mathbf{r}$. Consequently magnetic forces never cause a charge to speed up or slow down; they cause only a change in the charge's direction of motion.

Another consequence of this last property is that a uniform magnetic field perpendicular to a charge's plane of motion will result in a circular orbit, as

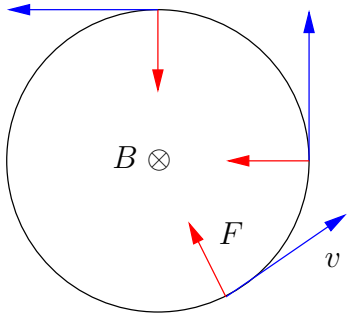


Figure 17.1: Magnetic Orbit of a Positive Charge with $\mathbf{B} \otimes$

shown in fig. (17.1) for the case that the magnetic field is into the page and a positive charge is moving in the plane of the page. In this case, the magnetic force (the red arrows in the figure) is not only always perpendicular to the charge's direction of motion (the blue arrows), but of constant magnitude: since the magnetic field is uniform, and since no magnetic force ever changes a charge's speed,

$$F = qvB \sin \frac{\pi}{2} = qvB = \text{const}$$

This is exactly the kind of force required for circular motion: a force always constant in magnitude and perpendicular to the velocity.

Magnetic fields are in fact used to keep charged particles like protons and antiprotons orbiting around the circular rings of particle colliders. And since the magnetic forces on charges of opposite sign will be in opposite directions, the same magnetic field that will hold protons in a clockwise orbit will hold antiprotons in a counterclockwise orbit — very convenient for smashing them together.

You should be able to work through any application of circular magnetic orbits by using the usual circular-motion force relation,

$$F_{magnetic} = F_{centripetal}$$

$$qvB = \frac{mv^2}{r} = m\omega^2 r$$

in conjunction with general relations for circular motion.

While we're at it, with a little gerrymandering we can also obtain a result for the force on a straight segment of current-carrying wire in a constant magnetic field: If a bit dq of the charge that constitutes the current undergoes a displacement $d\mathbf{r}$ down the wire during the time interval dt, the magnetic force on dq is

$$d\mathbf{F} = dq\, \mathbf{v} \times \mathbf{B} = dq\, \frac{d\mathbf{r}}{dt} \times \mathbf{B} = \frac{dq}{dt}\, d\mathbf{r} \times \mathbf{B} = I\, d\mathbf{r} \times \mathbf{B}$$

where $I = dq/dt$ is the current flowing down the wire. If we integrate this along the length ℓ of the wire, we can, since they are constant, pull the current I and field $\mathbf{B}$ outside the integration:

$$\mathbf{F} = \int_{\text{wire}} I\, d\mathbf{r} \times \mathbf{B} = I \left(\int_{\text{wire}} d\mathbf{r} \right) \times \mathbf{B}$$

The remaining integral over $d\mathbf{r}$ will yield simply the total displacement $\boldsymbol{\ell}$ of the charge along the wire, that is, a vector that points in the direction of current flow and has a magnitude equal to the length ℓ of the wire. In terms of this $\boldsymbol{\ell}$, for the force on the wire is

$$\mathbf{F} = I\boldsymbol{\ell} \times \mathbf{B} \tag{17.6}$$

17.2 Ampère's Law

Back to the magnetostatic versions (17.3) and (17.4) of Ampère's law,

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j} \qquad\qquad \oint_C \mathbf{B} \cdot d\mathbf{r} = \mu_0 I_{enclosed}$$

Note that the differential form of Ampère's law involves a cross product and thus a right-hand rule. The integral form, as you will recall from Stokes's theorem (2.8), involves a similar right-hand rule: we are free to choose the direction in which we perform the line integration around the loop C — known in this context as the Amperian loop — , but having made this choice, we must use the right-hand rule to determine which of the two directions for current flow through the loop is to be considered the positive direction: curling the fingers of your right hand around the loop in the direction in which you are integrating will leave your thumb pointed in the direction of positive current flow through the loop. Thus for a loop in the plane of the page, integrating around the loop counterclockwise would mean that current flowing out of the page should be considered positive; integrating clockwise that current flowing into the page should be considered positive.

Just as Gauss's law allowed us to determine the electric fields of highly symmetric distributions of charge, Ampère's law will allow us to determine the magnetic fields of highly symmetric distributions of current. *Highly symmetric* will, however, turn out to be much more restrictive in the case of Ampère's law: we will be able to determine the fields of only two kinds of current distributions: the infinite straight cylinder of current and the solenoid.[3]

17.2.1 Field of an Infinite Wire

Suppose a current I flows down an infinite cylindrical shell of radius a, uniformly distributed over its circular cross section. Cylindrical coordinates are most natural to the cylindrical symmetry of this distribution of current, and in cylindrical coordinates the most general expression for the magnetic field is

$$\mathbf{B}(\mathbf{r}) = B_r(r,\theta,z)\,\hat{\mathbf{r}} + B_\theta(r,\theta,z)\,\widehat{\boldsymbol{\theta}} + B_z(r,\theta,z)\,\hat{\mathbf{z}}$$

Because the distribution of current is symmetric under rotations by angles θ, the magnetic field, which must share this symmetry, cannot have any dependence on θ. And since shifting along an *infinite* cylinder cannot make any difference, neither can the field have any dependence on z. $\mathbf{B}$ can therefore depend only on r and reduces to

$$\mathbf{B}(\mathbf{r}) = B_r(r)\,\hat{\mathbf{r}} + B_\theta(r)\,\widehat{\boldsymbol{\theta}} + B_z(r)\,\hat{\mathbf{z}}$$

[3] For those of you unfamiliar with the term: a solenoid is essentially just a cylindrical coil of wire. We will deal with solenoids in §17.2.2.

Reasoning out the possible directions of the magnetic field is a bit more involved than it was for the electric field. First, we note that if the magnetic field were to emanate radially outward from a line, that would constitute a nonzero divergence that would contradict the magnetic Gauss's law $\nabla \cdot \mathbf{B} = 0$. Our magnetic field therefore cannot have a radial component. Second, we note from the cross product in the differential form

$$\nabla \times \mathbf{B} = \mu_0 \, \mathbf{j}$$

of Ampère's law that the magnetic field must always be perpendicular to the current density $\mathbf{j}$, that is, to the direction of current flow. Since our current flow is along the z axis, our magnetic field also cannot have any $\widehat{\mathbf{z}}$ component. So $\mathbf{B}$ further reduces to

$$\mathbf{B}(\mathbf{r}) = B_\theta(r) \, \widehat{\boldsymbol{\theta}} = B(r) \, \widehat{\boldsymbol{\theta}}$$

The Amperian loops C that reflect the symmetry of the current distribution are circles concentric with the cylindrical shell of current, in planes perpendicular to its axis. If we take $+\widehat{\mathbf{z}}$ as the direction of positive current flow, then the line element $d\mathbf{r}$ as we integrate around the loop will be $ds \, \widehat{\boldsymbol{\theta}}$, where $ds = r \, d\theta$ is the element of arc around the circle. Thus the left-hand side of the integral form of Ampère's law works out to

$$\begin{aligned} \oint_C \mathbf{B} \cdot d\mathbf{r} &= \oint_C B(r) \, \widehat{\boldsymbol{\theta}} \cdot r \, d\theta \, \widehat{\boldsymbol{\theta}} \\ &= \oint_C r B(r) \, d\theta \\ &= r B(r) \int_0^{2\pi} d\theta \\ &= r B(r) \, (2\pi) \\ &= 2\pi r B(r) \end{aligned}$$

where we have noted that since r is constant around our Amperian loop, $B(r)$ is also constant and can be brought outside the integral.

On the right-hand side of Ampère's law, how much current is enclosed by our Amperian loop depends on its radius: if the radius r of the Amperian loop is greater than the radius a of the shell of current, then all of the current I of the shell is enclosed; if the radius r of the Amperian loop is less than the radius a of the shell of current, then none of the current on the shell is enclosed. Ampère's law

$$\oint_C \mathbf{B} \cdot d\mathbf{r} = \mu_0 \, I_{enclosed}$$

therefore gives us

$$2\pi r B(r) = \begin{cases} \mu_0 I & (r > a) \\ \\ 0 & (r < a) \end{cases}$$

and hence

$$B(r) = \begin{cases} \dfrac{\mu_0 I}{2\pi r} & (r > a) \\ \\ 0 & (r < a) \end{cases} \tag{17.7}$$

A special case of (17.7) is an infinite straight wire of zero radius, for which eq. (17.7) reduces to

$$B = \frac{\mu_0 I}{2\pi r} \tag{17.8}$$

Another cylindrically symmetric distribution of current is a solid cylinder of radius a with a current I uniformly distributed over its cross section. The symmetry of this distribution, and hence the symmetry of the magnetic field and our result for the left-hand side of Ampère's law, will be the same as for the case of the cylindrical shell. The difference will be on the right-hand side of Ampère's law: while we again enclose the full current I of the cylinder when $r > a$, when $r < a$ we pick up only that fraction of the current I that lies within the radius r of our Amperian loop. Since the current is uniformly distributed over the cylinder's cross section, this fraction is proportional to the area enclosed:

$$I_{enclosed} = \frac{\pi r^2}{\pi a^2} I = \left(\frac{r}{a}\right)^2 I$$

So for the solid cylinder of current Ampère's law gives

$$2\pi r B(r) = \mu_0 \begin{cases} I & (r > a) \\ \\ \left(\dfrac{r}{a}\right)^2 I & (r < a) \end{cases}$$

and hence

$$B(r) = \begin{cases} \dfrac{\mu_0 I}{2\pi r} & (r > a) \\ \\ \dfrac{\mu_0 I}{2\pi}\,\dfrac{r}{a^2} & (r < a) \end{cases} \tag{17.9}$$

Remember that all of our above results for $B(r)$ are the $\widehat{\boldsymbol{\theta}}$ component and correspond to a current I in the $+\widehat{\mathbf{z}}$ direction. That our results for B come out positive indicates that for a current in the $+\widehat{\mathbf{z}}$ direction, we get a magnetic field $\mathbf{B}$ in the $+\widehat{\boldsymbol{\theta}}$ direction; a current in the $-\widehat{\mathbf{z}}$ would of course produce a field in the $-\widehat{\boldsymbol{\theta}}$ direction. Either way, the field is tangent to circles around the current; the issue is in which of the two possible tangential directions it points, and for that we now have another right-hand rule:

> If you point your thumb along the current, your fingers will curl around in the tangential direction of the magnetic field.

Note that if the current distribution in a cylinder is a function of r (so that, while still cylindrically symmetric, it is no longer uniform), then $I_{enclosed}$ is no longer proportional to the area enclosed; we have to perform the integration of eq. (14.6),

$$I = \int_S dA\, \widehat{\mathbf{n}} \cdot \mathbf{j}$$

If, for example, the current density within the cylinder went as $1/r$, then we would have

$$\mathbf{j} = C\frac{1}{r}\,\widehat{\mathbf{z}}$$

where the constant C would be determined by the condition that the total current flowing down the cylinder be I: for the circular disks S that constitute cross sections of the cylinder,

$$\begin{aligned} I &= \int_S dA\, \widehat{\mathbf{n}} \cdot \mathbf{j} \\ &= \int_S r\, dr\, d\theta\, \widehat{\mathbf{z}} \cdot C\frac{1}{r}\,\widehat{\mathbf{z}} \\ &= \int_S r\, dr\, d\theta\, C\frac{1}{r} \\ &= C\int_0^a dr \int_0^{2\pi} d\theta \\ &= C\,(a)\,(2\pi) \end{aligned}$$

which yields

$$C = \frac{I}{2\pi a}$$

and hence

$$\mathbf{j} = \frac{I}{2\pi a r}\,\widehat{\mathbf{z}}$$

So the current $I_{enclosed}$ enclosed by an Amperian loop of radius r would, by an almost identical integration, be

$$\begin{aligned} I_{enclosed} &= \int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{j} \\ &= \int_S r\, dr\, d\theta\, \hat{\mathbf{z}} \cdot \frac{I}{2\pi a r}\, \hat{\mathbf{z}} \\ &= \int_S r\, dr\, d\theta\, \frac{I}{2\pi a r} \\ &= \frac{I}{2\pi a} \int_0^r dr \int_0^{2\pi} d\theta \\ &= \frac{I}{2\pi a} (r)\, (2\pi) \\ &= I\, \frac{r}{a} \end{aligned}$$

You should be able to apply Ampère's law to the following cylindrically symmetric distributions of current:

- An infinite line.
- An infinite solid cylinder, of either uniform or nonuniform current density.
- An infinite cylindrical shell, of either infinitesimal or finite thickness.

17.2.2 Field of a Solenoid

A solenoid is essentially just a long coil of wire, what you would get by winding wire around a der-der tube,[4] as shown, less the der-der tube, in fig. (17.2). Not that fig. (17.2) is really going to clarify anything for you. But it's not like you don't already know what coils of wire look like, and fig. (17.2) probably would be cool to stare at when you're on drugs.

Consider now an infinitely long solenoid, as shown in cross section in fig. (17.3). The two series of little circles in fig. (17.3) represent the wire wound around the solenoid, through which we have sliced when we cut the solenoid lengthwise down its center. The colored rectangles in the figure are the Amperian loops to which we will later be applying Ampère's law. We will take the solenoid to be coaxial with the z axis, to be of radius a, and to

[4] A der-der tube is, of course, a paper-towel tube — the kind that, when the roll was empty, you used to run around the house with, holding it to your mouth and going "Der! Der! Der!" Ah, those were the days.

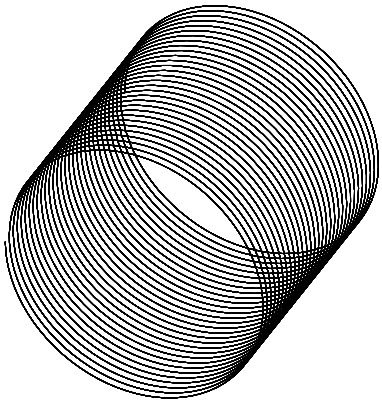


Figure 17.2: A Poor Man's Solenoid

have n turns of wire per unit length. The positive z axis is to the right, and a current I is flowing around the solenoid in such a way that the current is coming out at us in the top wires and going back into the page in the bottom wires.

Again there is a cylindrical symmetry, and our reasoning about the possible directions and dependencies of the magnetic field is very similar to that of the preceding section: the symmetry of the current distribution under rotations by angles θ and under shifts up or down the z axis rules out a θ or z dependence, and the requirement that the magnetic field have no divergence rules out a radial component. But whereas before the magnetic field had to be perpendicular to a current along the z axis, it must now be perpendicular to a current in the tangential (here specifically the $+\widehat{\boldsymbol{\theta}}$) direction and consequently cannot have a $\widehat{\boldsymbol{\theta}}$ component. So for our solenoid the general

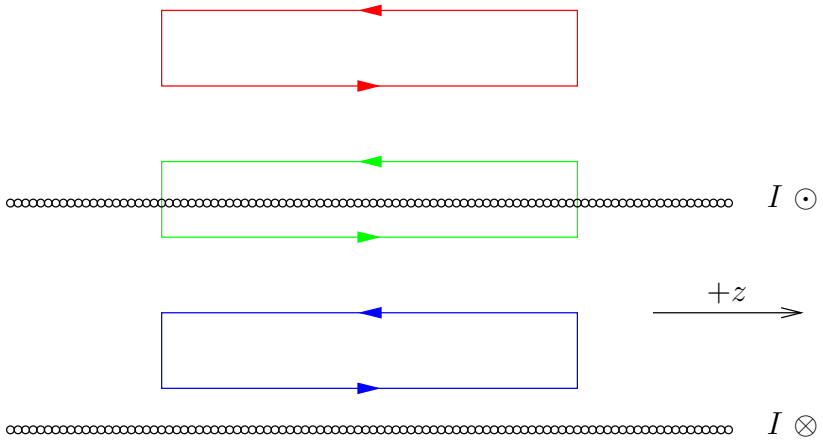


Figure 17.3: Anatomy of a Solenoid

expression

$$\mathbf{B}(\mathbf{r}) = B_r(r,\theta,z)\,\hat{\mathbf{r}} + B_\theta(r,\theta,z)\,\hat{\boldsymbol{\theta}} + B_z(r,\theta,z)\,\hat{\mathbf{z}}$$

for the magnetic field in cylindrical coordinates reduces to

$$\mathbf{B}(\mathbf{r}) = B_z(r)\,\hat{\mathbf{z}} = B(r)\,\hat{\mathbf{z}}$$

We now apply Ampère's law to the red Amperian loop shown in fig. (17.3). On the sides perpendicular to the axis of the solenoid (that is, perpendicular to the z axis), $d\mathbf{r}$ is perpendicular to $\hat{\mathbf{z}}$, so that $\mathbf{B}\cdot d\mathbf{r} = B(r)\,\hat{\mathbf{z}}\cdot d\mathbf{r} = 0$. So only the upper and lower sides of the loop make a nonvanishing contribution to the line integral of $\mathbf{B}$:

$$\oint_C \mathbf{B}\cdot d\mathbf{r} = \int\limits_{\text{lower side}} \mathbf{B}\cdot d\mathbf{r} \;+ \int\limits_{\text{upper side}} \mathbf{B}\cdot d\mathbf{r}$$

Since $d\mathbf{r} = \pm dz\,\hat{\mathbf{z}}$ along the lower and upper side, respectively, this becomes

$$\begin{aligned}\oint_C \mathbf{B}\cdot d\mathbf{r} &= \int\limits_{\text{lower side}} B(r)\,\hat{\mathbf{z}}\cdot dz\,\hat{\mathbf{z}} \;+ \int\limits_{\text{upper side}} B(r)\,\hat{\mathbf{z}}\cdot(-dz\,\hat{\mathbf{z}}) \\ &= \int\limits_{\text{lower side}} B(r)\,dz \;- \int\limits_{\text{upper side}} B(r)\,dz\end{aligned}$$

Because r is constant in value along each of these sides of the loop, so is $B(r)$, with $B = B(r_{lower})$ on the lower side and $B = B(r_{upper})$ on the upper side. Thus in terms of the length ℓ of these sides of the rectangle our line integral further simplifies to

$$\begin{aligned}\oint_C \mathbf{B}\cdot d\mathbf{r} &= B(r_{lower}) \int\limits_{\text{lower side}} dz \;- B(r_{upper}) \int\limits_{\text{upper side}} dz \\ &= B(r_{lower})\,\ell - B(r_{upper})\,\ell \\ &= \Big(B(r_{lower}) - B(r_{upper})\Big)\,\ell\end{aligned}$$

The result for the blue and green Amperian loops is identical, except of course that each loop has its own values of r_{lower} and r_{upper}.

No current is enclosed by the blue Amperian loop that is entirely inside the solenoid or the red Amperian loop that is entirely outside it, so for those two loops Ampère's law gives

$$\begin{aligned}\Big(B(r_{lower}) - B(r_{upper})\Big)\,\ell &= \mu_0 I_{enclosed} \\ &= 0\end{aligned}$$

and hence

$$B(r_{lower}) = B(r_{upper})$$

That is, the field is constant everywhere inside and everywhere outside the solenoid (though of course possibly differing in value between the inside and outside).

Since each wire carries the same current I and there are n turns of wire per unit length along the solenoid, the green Amperian loop that crosses the wires of the solenoid encloses a total current

$$I_{enclosed} = +In\ell$$

where the positive sign indicates that, since we are integrating around the loop counterclockwise, the positive direction for enclosed current is out of the page, which is in fact the direction in which the current I in the enclosed wires is flowing. Thus Ampère's law for this loop gives

$$\begin{aligned}\Big(B(r_{lower}) - B(r_{upper})\Big)\,\ell &= \mu_0 I_{enclosed} \\ &= \mu_0 In\ell\end{aligned}$$

and hence

$$B(r_{lower}) - B(r_{upper}) = \mu_0 nI \tag{17.10}$$

That is, the difference between the values of the magnetic field inside and outside the solenoid is $\mu_0 nI$.

To get an absolute result for the magnetic fields inside and outside the solenoid, we need more than the relative values that Ampère's law gives us. We need to know *the* value of the field at some point inside or outside the solenoid. And that point would be one infinitely far from the solenoid: although the solenoid is of infinite length, we have already seen that the field of an infinite wire falls off as $1/r$, and we expect the field of the solenoid to fall off even faster because of the cancellation from the opposite directions of the current flow at diametrically opposite points along its wall. We therefore expect field of the solenoid to vanish at locations infinitely far from it.[5] Since the field is constant everywhere outside the solenoid, it must therefore vanish everywhere outside, and consequently, by eq. (17.10), the constant field inside is

$$B = \mu_0 nI \tag{17.11}$$

Note that in our derivation the positive direction for magnetic field was to the right (the positive z direction) and the positive direction for current flow was as shown in fig. (17.3), with current coming out of the page over the top of the solenoid and going back into the page around the bottom (the $+\widehat{\boldsymbol{\theta}}$ direction). We thus arrive at yet another right-hand rule:

[5] We will establish this more rigorously when we do the solenoid by the Biot-Savart law in §17.3.5.

> If you wrap your fingers around the solenoid in the direction of the current flow, your thumb will point in the direction of its magnetic field.

Solenoids are the magnetic equivalent of parallel plates: they are a simple means of producing a field uniform in both magnitude and direction. But just as there were distortions of the electric field near the edges of finite parallel plates, there are distortions of the magnetic field near the ends of solenoids of finite length: the field is perfectly constant only for an infinitely long solenoid.

You should be able apply Ampère's law to solenoids. *Duh!*

17.3 The Biot-Savart Law

As in Bē-ō-Sə-vär′: Jean-Baptiste Biot and Félix Savart were French dudes.

17.3.1 Derivation of the Biot-Savart Law

Please fasten your seat belts. Our ultimate goal is a result for the magnetic field of a current I flowing down an arbitrarily curved wire, and the way there is a bit tortuous.

We start from the differential form of Ampère's law,

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{j}$$

In §2.7.1 we saw that the vanishing of the divergence of the magnetic field $\mathbf{B}$ ($\nabla \cdot \mathbf{B} = 0$) implied that $\mathbf{B}$ is a pure curl, that is, that $\mathbf{B}$ can be expressed in terms of a vector potential $\mathbf{A}$ as

$$\mathbf{B} = \nabla \times \mathbf{A}$$

Using this in Ampère's law, we have

$$\nabla \times (\nabla \times \mathbf{A}) = \mu_0 \mathbf{j} \tag{17.12}$$

As you can verify in a straightforward (if rather tedious) way simply by using

$$\nabla = \hat{\mathbf{x}} \frac{\partial}{\partial x} + \hat{\mathbf{y}} \frac{\partial}{\partial y} + \hat{\mathbf{z}} \frac{\partial}{\partial z}$$

to expand out the components of $\nabla \times (\nabla \times \mathbf{A})$, $\nabla(\nabla \cdot \mathbf{A})$, and $\nabla^2 \mathbf{A}$, the

relation [6]

$$\nabla \times (\nabla \times \mathbf{A}) = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A}$$

holds for any vector $\mathbf{A}$, so that we can re-express eq. (17.12) as

$$\nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} = \mu_0 \mathbf{j} \tag{17.13}$$

Inasmuch as this looks even more ornery than the $\nabla \times \mathbf{B} = \mu_0 \mathbf{j}$ that we started out with, it might not seem that we are making any progress. Recall from §14.8.1, however, that we can make the shift (known as a gauge transform)

$$\mathbf{A} \to \mathbf{A} + \nabla\Lambda$$

in the value of the vector potential $\mathbf{A}$, where Λ is an arbitrary function, without affecting the resulting physics in any way. This is easily verified for eq. (17.13):

$$\begin{aligned} \nabla\Big(\nabla \cdot (\mathbf{A} + \nabla\Lambda)\Big) - \nabla^2(\mathbf{A} + \nabla\Lambda) & \\ = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} + \nabla(\nabla^2\Lambda) - \nabla^2(\nabla\Lambda) & \\ = \nabla(\nabla \cdot \mathbf{A}) - \nabla^2 \mathbf{A} & \end{aligned}$$

This is very convenient, because by choosing a suitable function Λ we can make $\nabla \cdot \mathbf{A}$ vanish and simplify eq. (17.13) to [7]

$$-\nabla^2 \mathbf{A} = \mu_0 \mathbf{j}$$

[6] This is essentially just

$$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) = \mathbf{a} \cdot \mathbf{c}\, \mathbf{b} - \mathbf{a} \cdot \mathbf{b}\, \mathbf{c}$$

applied with $\mathbf{a} = \mathbf{b} = \nabla$ and care taken to keep the derivatives involved in ∇ to the left so that they act on $\mathbf{A}$.

[7] To see that this is always possible, recall from §2.7.3 that any vector field, including $\mathbf{A}$, can always be expressed as the sum of a pure gradient and a pure curl. Thus $\mathbf{A}$ must be of the form

$$\mathbf{A} = \nabla\alpha + \nabla \times \boldsymbol{\beta}$$

for some α and $\boldsymbol{\beta}$. All we need to do is choose $\Lambda = -\alpha$: then

$$\mathbf{A} \to \mathbf{A} + \nabla\Lambda = \nabla\alpha + \nabla \times \boldsymbol{\beta} - \nabla\alpha = \nabla \times \boldsymbol{\beta}$$

so that

$$\nabla \cdot \mathbf{A} = \nabla \cdot (\nabla \times \boldsymbol{\beta}) = 0$$

identically.

Making a choice of Λ to force $\mathbf{A}$ to satisfy a condition such as $\nabla \cdot \mathbf{A} = 0$ is called fixing a gauge. Since all physical results are independent of the choice of gauge, one is free to choose whatever gauge will make a calculation easiest. Our present choice $\nabla \cdot \mathbf{A} = 0$ is variously referred to as Coulomb gauge, transverse gauge, or radiation gauge.

Thus we arrive at

$$\nabla^2 \mathbf{A} = -\mu_0 \mathbf{j} \tag{17.14}$$

as an equivalent expression for Ampère's law in terms of the vector potential $\mathbf{A}$.

We can obtain a solution to eq. (17.14) by analogy to electrostatics: the general expression for the electric field in terms of the vector potential $\mathbf{A}$ and the scalar potential φ was given by eq. (14.19),

$$\mathbf{E} = -\nabla\varphi - \frac{\partial \mathbf{A}}{\partial t}$$

In electro- and magnetostatics, the fields have no time dependence, so that this reduces to

$$\mathbf{E} = -\nabla\varphi$$

If we use this together with Gauss's law

$$\nabla \cdot \mathbf{E} = \frac{1}{\epsilon_0}\,\rho$$

we have

$$\frac{1}{\epsilon_0}\,\rho = \nabla \cdot \mathbf{E} = \nabla \cdot (-\nabla\varphi) = -\nabla^2\varphi$$

Thus we arrive, as our equivalent expression for the electrostatic Gauss's law in terms of the scalar potential φ, at

$$\nabla^2\varphi = -\frac{1}{\epsilon_0}\,\rho \tag{17.15}$$

We already know the solution to this equation, in the sense that we already have solution (15.21) for φ:

$$\varphi = \frac{1}{4\pi\epsilon_0}\int \frac{dq}{r}$$

where "r" is the distance from the charge dq to where we are evaluating φ. Expressed more precisely, for a continuous distribution of charge density ρ, $dq = \rho(\mathbf{r}')\,dV'$ is the bit of charge located at $\mathbf{r}'$, and if we are evaluating φ at $\mathbf{r}$, the distance "r" is $|\mathbf{r} - \mathbf{r}'|$. Thus

$$\varphi = \frac{1}{4\pi\epsilon_0}\int \frac{\rho(\mathbf{r}')\,dV'}{|\mathbf{r} - \mathbf{r}'|} \tag{17.16}$$

Comparing eq. (17.15) to eq. (17.14), we see that we need simply make the correspondences

$$\varphi \leftrightarrow \mathbf{A} \qquad\qquad \frac{1}{\epsilon_0} \leftrightarrow \mu_0 \qquad\qquad \rho \leftrightarrow \mathbf{j}$$

By analogy to eq. (17.16), the solution to eq. (17.14) must therefore be

$$\mathbf{A} = \frac{\mu_0}{4\pi} \int \frac{\mathbf{j}(\mathbf{r}')\, dV'}{|\mathbf{r} - \mathbf{r}'|} \tag{17.17}$$

We have now succeeded in solving for the vector potential $\mathbf{A}$ in terms of the current density $\mathbf{j}$. To get a result for $\mathbf{B} = \nabla \times \mathbf{A}$, we need to take the curl of both sides of eq. (17.17) — tedious, but fairly straightforward. First, note that the $\nabla\times$ involves derivatives with respect to (x, y, z), not (x', y', z'). That is, $\nabla\times$ will act on

$$\mathbf{r} = x\,\widehat{\mathbf{x}} + y\,\widehat{\mathbf{y}} + z\,\widehat{\mathbf{z}}$$

and not on

$$\mathbf{r}' = x'\,\widehat{\mathbf{x}} + y'\,\widehat{\mathbf{y}} + z'\,\widehat{\mathbf{z}}$$

Since the only dependence on $\mathbf{r}$ is in the $|\mathbf{r} - \mathbf{r}'|$, we thus have

$$\begin{aligned}
\mathbf{B} &= \nabla \times \mathbf{A} \\
&= \nabla \times \frac{\mu_0}{4\pi} \int dV' \, \frac{\mathbf{j}(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \\
&= \frac{\mu_0}{4\pi} \int dV' \left(\nabla \frac{1}{|\mathbf{r} - \mathbf{r}'|} \right) \times \mathbf{j}(\mathbf{r}')
\end{aligned} \tag{17.18}$$

Now, the x component of

$$\nabla \frac{1}{|\mathbf{r} - \mathbf{r}'|}$$

works out by the chain rule to

$$\begin{aligned}
&\frac{\partial}{\partial x} \frac{1}{\sqrt{(x-x')^2 + (y-y')^2 + (z-z')^2}} \\
&\qquad = -\frac{1}{2} \frac{1}{\Big((x-x')^2 + (y-y')^2 + (z-z')^2\Big)^{\frac{3}{2}}} \Big(2(x-x')\Big) \\
&\qquad = -\frac{x - x'}{\Big((x-x')^2 + (y-y')^2 + (z-z')^2\Big)^{\frac{3}{2}}}
\end{aligned}$$

and likewise for the y and z components, so that

$$\nabla \frac{1}{|\mathbf{r} - \mathbf{r}'|} = -\frac{(x-x')\,\widehat{\mathbf{x}} + (y-y')\,\widehat{\mathbf{y}} + (z-z')\,\widehat{\mathbf{z}}}{\Big((x-x')^2 + (y-y')^2 + (z-z')^2\Big)^{\frac{3}{2}}} = -\frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3}$$

Eq. (17.18) thus becomes

$$\mathbf{B} = \frac{\mu_0}{4\pi} \int dV' \left(-\frac{\mathbf{r} - \mathbf{r}'}{|\mathbf{r} - \mathbf{r}'|^3} \right) \times \mathbf{j}(\mathbf{r}')$$

$$= -\frac{\mu_0}{4\pi}\int dV'\,\frac{(\mathbf{r}-\mathbf{r}')\times\mathbf{j}(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3}$$

$$= \frac{\mu_0}{4\pi}\int dV'\,\frac{\mathbf{j}(\mathbf{r}')\times(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} \tag{17.19}$$

where we have in the last step absorbed the negative sign by reversing the order of the cross product.

Eq. (17.19) is a general result for the magnetic field due to a distribution $\mathbf{j}$ of current density. For the special case of a current I flowing through a wire, we have, by eqq. (14.5) and (14.4),

$$dV'\,\mathbf{j}(\mathbf{r}') = dV'\,\rho(\mathbf{r}')\mathbf{v}' = dV'\,\frac{dq'}{dV'}\,\frac{d\mathbf{r}'}{dt} = \frac{dq'}{dt}\,d\mathbf{r}' = I\,d\mathbf{r}' \tag{17.20}$$

where $d\mathbf{r}'$ is the displacement the charge dq' undergoes in time dt — in other words, $d\mathbf{r}'$ is a segment of the wire that, as a vector, points in the direction of current flow. Using eq. (17.20) in eq. (17.19) gives

$$\mathbf{B} = \frac{\mu_0 I}{4\pi}\int \frac{d\mathbf{r}'\times(\mathbf{r}-\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|^3} \tag{17.21}$$

This can be written in a form that is easier to remember: note that $\mathbf{r}$ and $\mathbf{r}'$ occur only in the combination $\mathbf{r}-\mathbf{r}'$, which, as shown in fig. (17.4), points from the current traveling through the segment $d\mathbf{r}'$ to the location $\mathbf{r}$ where we are evaluating the magnetic field. We will therefore simplify our notation by writing $\mathbf{R}$ in place of $\mathbf{r}-\mathbf{r}'$ and $d\mathbf{s}$ in place of $d\mathbf{r}'$:

$$\mathbf{B} = \frac{\mu_0 I}{4\pi}\int \frac{d\mathbf{s}\times\mathbf{R}}{R^3} \tag{17.22}$$

Eq. (17.22) is the form of the Biot-Savart law that we will be using.

When applying eq. (17.22), you have two tasks: to evaluate the direction given by the cross product $d\mathbf{s}\times\mathbf{R}$ and then to carry out the integration. Remember that you are integrating over the segments $d\mathbf{s}$ that make up the wire, that $d\mathbf{s}$ points in the direction of the current I, and that $\mathbf{R}$ is the vector *from the current* to where you are evaluating the magnetic field.

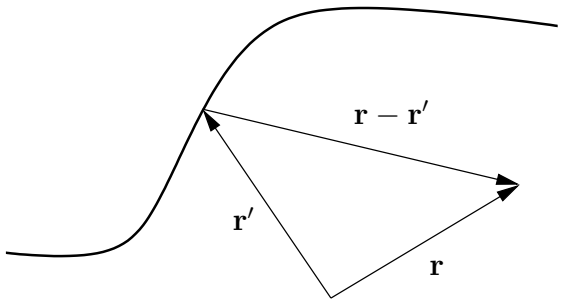


Figure 17.4: $\mathbf{r}$ Versus $\mathbf{r}'$

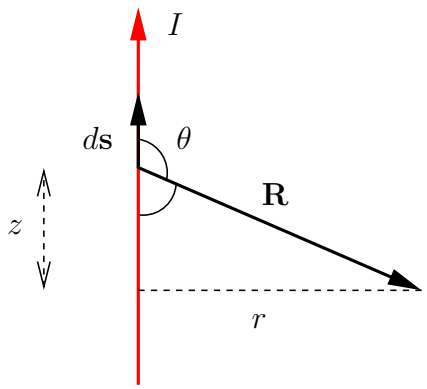


Figure 17.5: The Return of the Infinite Straight Wire

17.3.2 Field of an Infinite Wire

Suppose we have an infinite straight wire carrying current I and want to determine the magnetic field at a perpendicular distance r from the wire, as shown in fig. (17.5), in which the red arrow represents the current. The cross product $d\mathbf{s} \times \mathbf{R}$ is into the page and involves the sine of the angle θ between $d\mathbf{s}$ and $\mathbf{R}$. The sine of this angle is the same as the sine of its supplementary angle, which, as you can see from the figure, is r/R. Thus

$$\begin{aligned} d\mathbf{s} \times \mathbf{R} &= ds\, R \sin\theta \otimes \\ &= ds\, R \frac{r}{R} \otimes \\ &= r\, ds \otimes \end{aligned}$$

If we make the wire our z axis, with the origin at the base of the dashed line, $ds = dz$ and

$$R = \sqrt{r^2 + z^2}$$

The Biot-Savart law then works out to

$$\begin{aligned} \mathbf{B} &= \frac{\mu_0 I}{4\pi} \int_{\text{wire}} \frac{d\mathbf{s} \times \mathbf{R}}{R^3} \\ &= \frac{\mu_0 I}{4\pi} \int_{\text{wire}} \frac{r\, ds}{R^3} \otimes \\ &= \frac{\mu_0 I}{4\pi} \int_{-\infty}^{\infty} \frac{r\, dz}{(r^2 + z^2)^{\frac{3}{2}}} \otimes \\ &= \frac{\mu_0 I}{4\pi}\, 2 \int_{0}^{\infty} \frac{r\, dz}{(r^2 + z^2)^{\frac{3}{2}}} \otimes \end{aligned}$$

This is the familiar [8] tangent-substitution integral, which yields

$$\mathbf{B} = \frac{\mu_0 I}{4\pi}\, 2\,\frac{1}{r} \otimes = \frac{\mu_0 I}{2\pi r} \otimes$$

This is of course the same $B = \mu_0 I/2\pi r$ that we got from Ampère's law, with the direction of $\mathbf{B}$ given by the same right-hand rule.

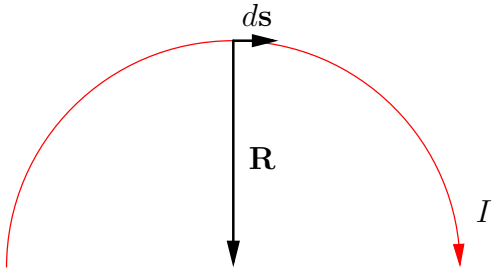


Figure 17.6: An Arc of Current

17.3.3 Field of a Circular Arc

Now suppose we are at the center of a circular arc of radius R that carries a current I, as shown in fig. (17.6).[9] Since there is always a right angle between $d\mathbf{s}$ and $\mathbf{R}$,

$$\begin{aligned} d\mathbf{s} \times \mathbf{R} &= ds\, R \sin\frac{\pi}{2} \otimes \\ &= R\, ds \otimes \end{aligned}$$

And since R is constant around the arc, the Biot-Savart law works out to

$$\begin{aligned} \mathbf{B} &= \frac{\mu_0 I}{4\pi} \int_{\text{arc}} \frac{d\mathbf{s} \times \mathbf{R}}{R^3} \\ &= \frac{\mu_0 I}{4\pi} \int_{\text{arc}} \frac{R\, ds}{R^3} \otimes \\ &= \frac{\mu_0 I}{4\pi} \frac{1}{R^2} \int_{\text{arc}} ds \otimes \end{aligned}$$

[8] Familiar as in p.792.

[9] While this arc does not form a closed loop and therefore does not constitute a magnetostatic current, it could be merely part of a more complex configuration of current that does form a closed loop, and the Biot-Savart law can certainly be applied to this arc's contribution to the net magnetic field.

$$= \frac{\mu_0 I}{4\pi} \frac{1}{R^2} (R\,\Delta\theta) \otimes$$

$$= \frac{\mu_0 I}{4\pi R} \Delta\theta \otimes$$

where $\Delta\theta$ is the angular extent of the arc (π for a semicircle, etc.). For a full circle, $\Delta\theta = 2\pi$ and

$$B = \frac{\mu_0 I}{2R}$$

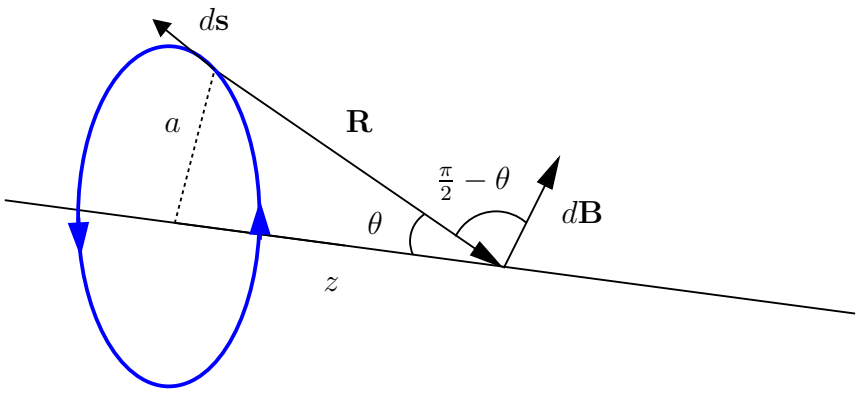


Figure 17.7: A Ring of Current

17.3.4 Field of a Ring of Current

Next, suppose we have a ring of radius a carrying a current I and want to determine the magnetic field at points on the axis of the ring, as "illustrated" in fig. (17.7), which attempts to show, from an even more cryptically medieval perspective than usual, a current flowing around the ring such that it is coming obliquely out of the page at the top of the ring and going obliquely into the page at the bottom of the ring. Thus the greatly exaggerated $d\mathbf{s}$ toward the top of the ring is coming obliquely out of page.

The first thing to note is that although there are now some funky angles involved, no matter where we are on the ring the angle between $d\mathbf{s}$ and $\mathbf{R}$ is always a right angle. So the magnitude of $d\mathbf{s} \times \mathbf{R}$ is

$$|d\mathbf{s} \times \mathbf{R}| = ds\, R \sin\frac{\pi}{2} = R\, ds$$

and its direction is that of the contribution $d\mathbf{B}$ to the net magnetic field shown in fig. (17.7). As we integrate around the ring, these contributions $d\mathbf{B}$ will trace out a cone, so that only the components along the ring's axis will

survive when they are added as vectors. The trig factor that extracts these components along the ring's axis is

$$\cos\left(\frac{\pi}{2}-\theta\right)=\sin\theta=\frac{a}{R}$$

And if we make the ring's axis the z axis, with the origin at the center of the ring, then R, which is constant around the ring, is given by

$$R=\sqrt{a^2+z^2}$$

If we put all this together, the Biot-Savart law works out to

$$\begin{aligned}
\mathbf{B} &= \frac{\mu_0 I}{4\pi}\int_{\text{ring}}\frac{d\mathbf{s}\times\mathbf{R}}{R^3} \\
&= \frac{\mu_0 I}{4\pi}\int_{\text{ring}}\frac{|d\mathbf{s}\times\mathbf{R}|}{R^3}\ (\text{trig factor})\ \hat{\mathbf{z}} \\
&= \frac{\mu_0 I}{4\pi}\int_{\text{ring}}\frac{R\,ds}{R^3}\,\frac{a}{R}\,\hat{\mathbf{z}} \\
&= \frac{\mu_0 I}{4\pi}\frac{a}{R^3}\int_{\text{ring}} ds\,\hat{\mathbf{z}} \\
&= \frac{\mu_0 I}{4\pi}\frac{a}{(a^2+z^2)^{\frac{3}{2}}}\,(2\pi a)\,\hat{\mathbf{z}} \\
&= \frac{\mu_0 I}{2}\,\frac{a^2}{(a^2+z^2)^{\frac{3}{2}}}\,\hat{\mathbf{z}} \qquad (17.23)
\end{aligned}$$

When we are at the center of the ring ($z=0$), this reduces to

$$B=\frac{\mu_0 I}{2a}$$

in agreement with our result for a full circular arc in the preceding section.

17.3.5 Field of a Solenoid

To get a result for the magnetic field of an infinite solenoid at points along its axis, all we have to do is integrate the fields (17.23) of a succession of rings extending from $z=-\infty$ to $z=+\infty$: if there are n rings (turns) per unit length, the number of rings in the interval dz is $n\,dz$, so that, with a contribution of the form (17.23) from each of the $n\,dz$ rings, we have a net magnetic field

$$B=\int_{-\infty}^{\infty} n\,dz\,\frac{\mu_0 I}{2}\,\frac{a^2}{(a^2+z^2)^{\frac{3}{2}}}=\frac{\mu_0 n I}{2}\,2\int_0^{\infty} dz\,\frac{a^2}{(a^2+z^2)^{\frac{3}{2}}}$$

This is once again the familiar tangent-substitution integral and yields

$$B = \frac{\mu_0 nI}{2}\, 2\,(1) = \mu_0 nI$$

in agreement with eq. (17.11).

You should be able to use the Biot-Savart law to determine the magnetic field of the following distributions of current:

- An infinite straight line of current.
- A circular arc of current, at the center of the circle.
- A ring of current, at points along the axis of the ring.

17.4 Magnetic Dipoles

A magnetic dipole in its simplest form consists of a little loop of current, which for simplicity we will in our analysis take to be a square of side ℓ carrying current I. Fig. (17.8) shows a top view of the loop, with its current circulating counterclockwise and a constant external magnetic field B — as opposed to the magnetic field of the loop itself, about which we are not worried here — coming out of the page. Also shown, in blue, are the directions of the magnetic forces $\mathbf{F} = I\boldsymbol{\ell} \times \mathbf{B}$ exerted by the external field on each of the four sides of the loop. Fig. (17.9) shows the same loop from a side view: the thick black line is the edge of the loop, which is now revealed to be tilted at angle θ. Perhaps contrary to your initial impression, in fig. (17.8) the loop is therefore not actually in the plane of the page. But then we never said

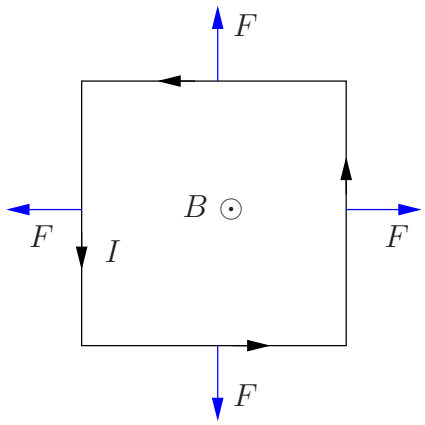


Figure 17.8: A Magnetic Dipole

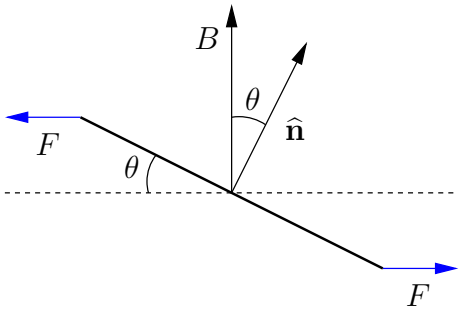


Figure 17.9: Déjà Vu All Over Again

it was, and you should know better than to make judgements based on first impressions. Anyway, in fig. (17.9) $\hat{\mathbf{n}}$ is the normal to the loop.

Consider first the top and bottom of the loop in fig. (17.8) — the sides we are looking at edge-on in fig. (17.9). The magnitude of the magnetic forces on these two sides of the loop is

$$F = |I\boldsymbol{\ell} \times \mathbf{B}| = I\ell B_\perp = I\ell B \cos\theta$$

These two forces, being opposite in direction, make no contribution to the net force. They also make no contribution to the net torque about the center of the loop, since they have zero torque arm about that point.

Now consider the left and right sides of the loop in fig. (17.8) — the sides we are looking down the ends of in fig. (17.9). The magnitude of the magnetic forces on these two sides of the loop is

$$F = |I\boldsymbol{\ell} \times \mathbf{B}| = I\ell B_\perp = I\ell B$$

Since these forces are again opposite in direction, they too make no contribution to the net force. There is therefore no net force on the loop. As you can see from fig. (17.10), this pair of forces does, however, make a net contribution to the net torque about the center of the loop, since they each have nonzero torque arm

$$r_\perp = \tfrac{1}{2}\ell \sin\theta$$

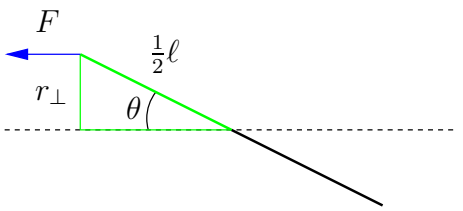


Figure 17.10: The Torque Arms

and both tend to rotate the loop counterclockwise from the perspective of fig. (17.9). The net counterclockwise torque on the loop is therefore

$$\tau = 2r_\perp F = 2(\tfrac{1}{2}\ell \sin\theta)(I\ell B) = I\ell^2 B \sin\theta = IAB\sin\theta$$

where $A = \ell^2$ is the area of the loop.

As a vector, this counterclockwise net torque points out of the page from the perspective of fig. (17.9). Now, the normal $\hat{\mathbf{n}}$ to the loop show in that figure is in a right-handed sense relative to the direction of the current circulation around the loop, and as you can see from the figure

$$\hat{\mathbf{n}} \times \mathbf{B} = B\sin\theta \odot$$

We can therefore write the net torque on the loop as

$$\boldsymbol{\tau} = IA\,\hat{\mathbf{n}} \times \mathbf{B}$$

or

$$\boldsymbol{\tau} = \boldsymbol{\mu} \times \mathbf{B} \tag{17.24}$$

if we define the magnetic dipole moment $\boldsymbol{\mu}$ of the loop to be

$$\boldsymbol{\mu} = IA\,\hat{\mathbf{n}} \tag{17.25}$$

Although we have derived eq. (17.24) only for a square loop at a simple tilt to the external magnetic field, it is actually a general relation, independent of the geometry or nature of the dipole. And eq. (17.25) similarly turns out to apply to planar loops of any shape, not just squares.

Note in particular that, as you can see by applying the right-hand rule to the current loop in fig. (17.8), a magnetic dipole's own field is in the same direction as its dipole moment $\boldsymbol{\mu}$. This is the reverse of the situation with an electric dipole, the dipole moment $\mathbf{p}$ of which points from the negative to the positive charge and is thus opposite in direction to the dipole's own field.

To get a result for the potential energy of a magnetic dipole in an external magnetic field, recall eq. (7.27) for the work done by a torque:

$$W = \int \boldsymbol{\tau} \cdot d\boldsymbol{\theta}$$

If we use eq. (17.24) and measure our angles $\boldsymbol{\theta}$ along the same axis as the torque $\boldsymbol{\tau}$, with $\theta = 0$ aligned with $\mathbf{B}$ so that θ coincides with the angle between $\boldsymbol{\mu}$ and $\mathbf{B}$, then the work done on the dipole by the external magnetic field as we rotate the dipole from angle θ_0 to angle θ is

$$\begin{aligned} W &= \int_{\theta_0}^{\theta} \tau\, d\theta \\ &= \int_{\theta_0}^{\theta} |\boldsymbol{\mu} \times \mathbf{B}|\, d\theta \end{aligned}$$

$$
\begin{aligned}
&= \int_{\theta_0}^{\theta} \mu B \sin\theta \, d\theta \\
&= \Big[-\mu B \cos\theta \Big]_{\theta_0}^{\theta} \\
&= -\mu B \cos\theta + \mu B \cos\theta_0
\end{aligned}
$$

Since this work done on the dipole can be regarded as stored in the dipole, we have as the dipole's potential energy

$$U = -\mu B \cos\theta + \mu B \cos\theta_0$$

If for simplicity we choose $\theta_0 = \frac{\pi}{2}$, this reduces to

$$U = -\mu B \cos\theta$$

which, since θ is the angle between the field and the dipole, may be more neatly expressed as

$$U = -\boldsymbol{\mu} \cdot \mathbf{B} \tag{17.26}$$

Note that, like the electric dipoles of §15.5, our magnetic dipoles are presumed to be small in size, so small that even if the external magnetic field is not (as we have assumed throughout this section) constant in magnitude and direction, it will not vary significantly over the dipole. The dipoles of interest are in fact often atomic or molecular, with the orbits of a unpaired outer electrons constituting the loops of current. The behavior of such magnetic dipoles in an external magnetic field is very similar to that of electric dipoles in an electric field: Because there is no net force on the dipoles, they do not move in response to the field, but they do rotate: to the extent that a molecule's dipole moment $\boldsymbol{\mu}$ is not aligned with the $\mathbf{B}$ field, there is a torque on it, and from $U = -\boldsymbol{\mu} \cdot \mathbf{B}$ we see that the lowest energy state is when the dipole moment $\boldsymbol{\mu}$ points in the same direction as $\mathbf{B}$. Thus molecular dipoles tend, modulo the random jostling of thermal motion, to align with an external magnetic field. But whereas the field of an electric dipole is opposite in direction to the external field and thus diminishes the net electric field, the field of a magnetic dipole is in the same direction as the external field with which it is aligning and thus augments the net magnetic field. This is the cause of paramagnetism, the tendency of certain substances to become magnetic in the presence of an external field.

Ferromagnetism is an extreme case of paramagnetism: once they have been aligned by an external field, the field of the dipoles of a ferromagnet is strong enough to hold them in alignment by itself. The result is a permanent magnet that persists even after the external field is removed. In fact, even in the absence of an external field the dipoles will tend to align with each other; it is just that when there is no external field to "guide" them, spontaneous

alignments along random directions occur at many points within the substance, with the result that myriad domains are formed, each with its own random direction for its net magnetic field, and the net field over macroscopic regions of the substance remains zero.

To make a permanent magnet out of, say, a lump of iron, you first heat the iron so that thermal jostling thoroughly randomizes the orientations of the dipoles. You then impose an external field, with which the dipoles will freeze into alignment as the iron cools, and *voilà*, you've got a permanent magnet. This happens naturally to iron that solidifies in the lava emerging from the Mid-Atlantic Ridge, and the permanent magnets so produced record the history of the Earth's magnetic field: the separating plates act as conveyor belts carrying the fresh rock away from the ridge; the farther from the ridge, the older the rock. And it turns out the history of the Earth's magnetic field is not as dull as you might have thought: for reasons still unknown, it suddenly [10] reverses itself at irregular intervals on the order of anywhere from 10^4 to 10^6 yr.

Finally, although we will not be getting into it — or even doing much with magnetic dipoles, for that matter — , we should note that, like the electric field, the magnetic field has a multipole expansion: there are magnetic dipoles, quadrupoles, octopoles, etc., the field lines of which are identical to those of their electric kindred illustrated in Appendix F.

[10] In the geologic sense of suddenly, at any rate. Even if such a reversal were to occur during your lifetime, it's not as though you'd see migrating birds doing a 180 in mid flight.

17.5 Problems

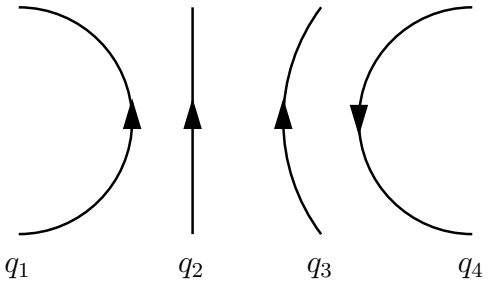


Figure 17.11: Problem 1

1. Four charges of equal mass, q_1, q_2, q_3, and q_4, are moving along the trajectories shown in fig. (17.11) in the directions indicated by the arrows. Everywhere within the region shown, the magnetic field is constant and out of the page ($\odot$). What can you conclude about the sign and relative magnitude of each charge?

2. At some instant, a positive charge q has the velocity vector

$$\mathbf{v} = v_x\,\hat{\mathbf{x}} + v_y\,\hat{\mathbf{y}} + v_z\,\hat{\mathbf{z}}$$

in a region where the magnetic field is $\mathbf{B} = B_0\,\hat{\mathbf{z}}$, where B_0 is a constant. Describe the trajectory of the charge — the more quantitatively, the better.

3. A mass m carrying a positive charge q is moving at some instant with velocity $\mathbf{v} = v\,\hat{\mathbf{x}}$. If the magnetic force experienced by the mass at that moment is $\mathbf{F}_{mag} = F_{mag}\,\hat{\mathbf{y}}$, what can you conclude about the magnitude, direction, and components of the magnetic field at the mass's location?

4. Determine the angular frequency (known as the cyclotron frequency) of a charge q of mass m orbiting in a circle of radius r in a region of constant magnetic field B.[11]

[11] What the **** is a cyclotron? For right now, you'll have to make do with the rather shallow answer that it was an early kind of particle accelerator; a more detailed description of this animal, and discussion of the practical consequences your result for the cyclotron frequency, will have to wait until class.

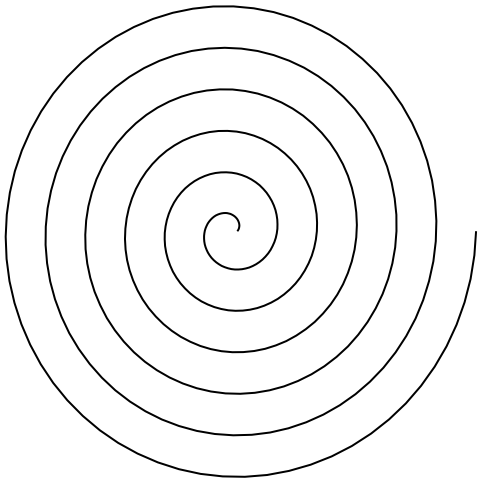


Figure 17.12: Problem 5

5. Back in the old days, before there were more sophisticated detectors, the paths of particles produced by particle colliders were observed by means of a cloud chamber: as the charged particles traveled under the influence of a magnetic force through a supercooled vapor, they left a trail of ions that in turn led to condensation of the vapor, creating what looked like the "con trails" of a miniature jet planes. Measurements were later made by hand on photographs of these trails.

   One commonly observed trail was like the inward spiral shown in fig. (17.12). If the cloud chamber was immersed in a constant magnetic field directed out of the page ($\odot$) and of known magnitude, and if you could make any distance measurements you wanted on the photograph of the trajectory, what could you conclude about the sign and magnitude of the particle's charge? About its mass? About its velocity? Why is the trajectory an inward spiral rather than a circle? See the footnote if you need a hint.[12]

[12] The particle will be losing energy as it collides with and ionizes vapor molecules.

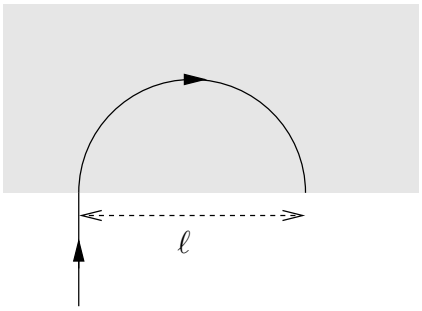


Figure 17.13: Problem 6

6. Fig. (17.13) shows the bare basics of a mass spectrometer, a device for separating ionized molecules or atoms according to their mass:[13] The ions, which we will suppose all to have charge q and mass m, are collimated into a very narrow beam after being accelerated through a potential difference V_0. This beam then enters a region, indicated in gray in fig. (17.13), of constant magnetic field B perpendicular to the plane of the page. Determine the distance ℓ from the beam's entry point at which the collecting "bucket" should be placed.

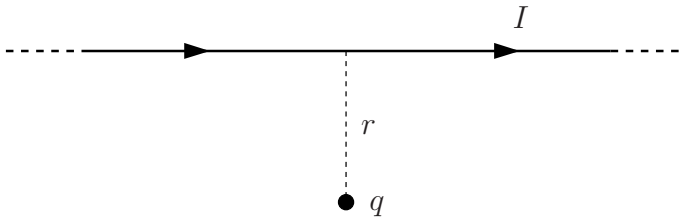


Figure 17.14: Problem 7

7. Fig. (17.14) shows a negative point charge q with no real sense of purpose in life just drifting around near an infinite straight line of current I. At the dramatic moment depicted in the figure, the perpendicular distance between the charge and the wire is r. Determine the magnitude and direction of the magnetic force experienced by the point charge when it is moving at speed v

   (a) Perpendicularly toward the wire.
   (b) Perpendicularly away from the wire.
   (c) In the same direction as the current.
   (d) Into the page ($\otimes$).
   (e) To the upper right at angle θ to the figure's horizontal direction.

[13] Mass spectrometers can be used to separate small amounts of the various isotopes of an element in a sample, but they aren't practicable for large jobs, such as separating enough uranium-235 to make a bomb. Just in case you had any ideas along those lines.

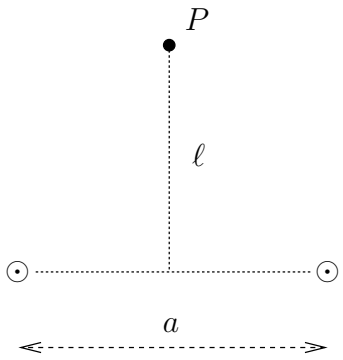


Figure 17.15: Problem 8

8. The two $\odot$ symbols in fig. (17.15) represent an end view of two wires, each carrying a current I out of the page.

   (a) Isn't this exciting?

   (b) Determine the magnitude and direction of the net magnetic field at point P, which is an equal distance from the two wires.

   (c) Show that this result for the net magnetic field reduces to what you would expect when

      i. $\ell \to 0$.

      ii. ℓ is large.

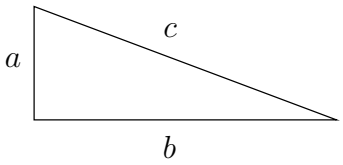


Figure 17.16: Problem 9

9. A counterclockwise current I flows around the perimeter of the right triangle in fig. (17.16). In the neighborhood where the triangle lives, there is a constant magnetic field B directed out of the page.

   (a) Determine the magnetic force on each of the three sides of the triangle explicitly and show that the net magnetic force on the triangle vanishes.

   (b) Show that quite generally the net magnetic force on any loop, whether planar or not, vanishes when the loop is in a region of constant magnetic field. See the footnote if you need a hint.[14]

[14] Integrate the infinitesimal contributions to the net force around the loop.

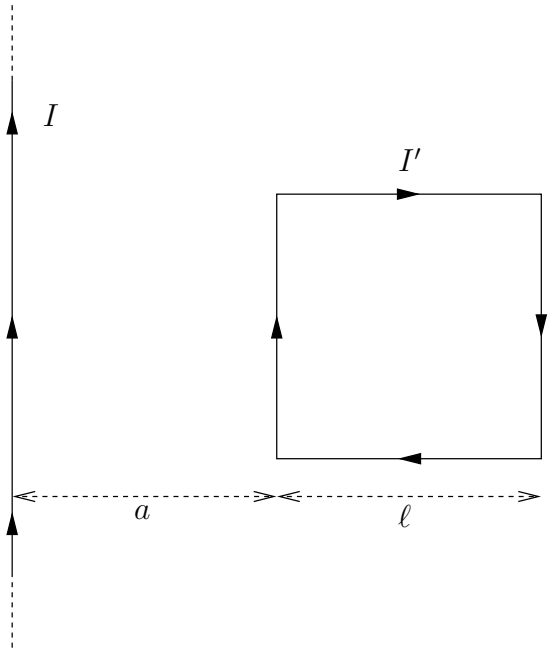


Figure 17.17: Problem 10

10. Fig. (17.17) shows an infinite straight wire carrying a current I up the page. To add to the excitement, a square of side ℓ, lying a perpendicular distance a from the wire and in a plane with it, carries a clockwise current I'. Determine the magnitude and direction of the net magnetic force exerted on the square by the infinite wire.

11. (a) Show that parallel wires carrying currents I_1 and I_2 exert on each other a magnetic force per unit length F/ℓ given by

$$\frac{F}{\ell} = \frac{\mu_0 I_1 I_2}{2\pi r}$$

where r is the perpendicular distance between the wires.

(b) How does the direction of this force depend on the directions of the currents carried by the wires?

(This illustrates how electromagnetism is neither right- nor left-handed and thus respects a spacetime symmetry known as parity,[15] even though the relations for magnetic fields and forces both involve cross products evaluated by right-hand rules. The key is that ultimately a magnetic force always involves *two* right-hand rules, one in the relation for the magnetic force, and another in the relation for the magnetic field that gives rise to that magnetic force: as the case of parallel wires shows, two right-hand rules in succession lead to a result that is neither right- nor left-handed. Just as we could equally well have chosen the opposite sign convention for electric charge, taking electrons to be positive and nuclei to be negative, using left-hand rules rather than right-hand rules in electromagnetism would yield the same physics.)

12. Explain why you can't use Ampère's law to determine the magnetic field of a current I flowing down a wire of square cross section, even if the current is evenly distributed over its cross section.

13. What choice of Amperian loop would allow you to determine the magnetic field of a current I flowing around the perimeter of a circle of radius a by Ampère's law?

[15] We will discuss parity in Chapter 22. You can find a basic description of it on p.1072.

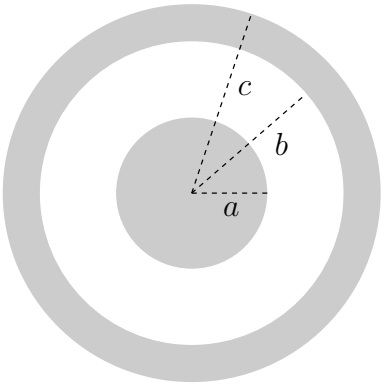


Figure 17.18: Problem 14

14. Fig. (17.18) shows a cross section of a coaxial cable: a current I travels down an inner wire of radius a and back up an outer "wire" braided into the form of a coaxial cylinder of inner radius b and outer radius c. We will assume that the current I is evenly distributed over the cross sections of the inner and outer wires and that it is into the page in the inner wire and out of the page in the outer wire.

    (a) Determine the net magnetic field everywhere (that is, for all r: $r \leq a$, $a \leq r \leq b$, $b \leq r \leq c$, and $r \geq c$).

    (b) What practical advantage does coaxial cable therefore have over ordinary wires, in which the current travels up and down two wires that run side by side?

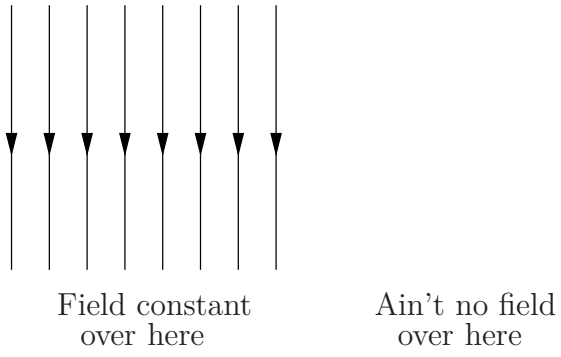


Figure 17.19: Problem 15

15. (Ye olde fringe-field problem.) Use Ampère's law to show that in a current-free region you cannot pass suddenly from a constant magnetic field strength to zero magnetic field, as shown in fig. (17.19).

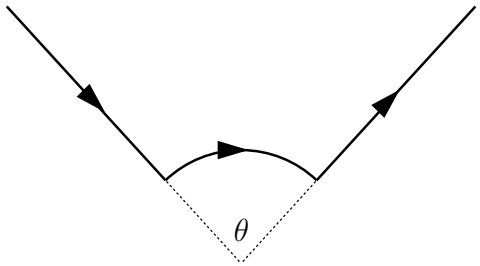


Figure 17.20: Problem 16

16. Fig. (17.20) shows a current I flowing radially inward from infinitely far away, around an arc of angle θ and radius a, and then radially back out to infinity. Determine the net magnetic field at the center of the arc.

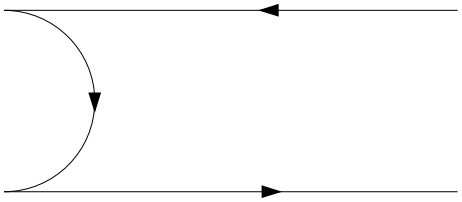


Figure 17.21: Problem 17

17. Fig. (17.21) shows a current I flowing inward from infinitely far away along a tangent to a semicircular arc of radius a, around the arc, and then back out to infinity along another tangent. Determine the net magnetic field at the center of the semicircular arc.

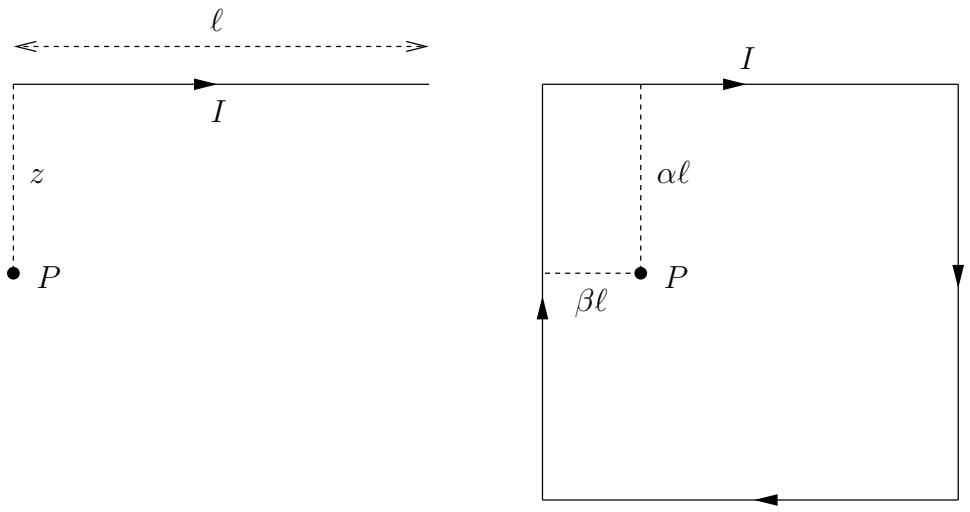


Figure 17.22: Problem 18

18. (a) The left side of fig. (17.22) shows a line segment of length ℓ that carries a current I.[16] Determine the magnitude and direction of the magnetic field at the point P, which lies a perpendicular distance z from one end of the segment.

    (b) The right side of fig. (17.22) shows a general point P inside a square of side ℓ around which flows a clockwise current I. We will parametrize the location of P by α and β, where the distance of the point P from the nearest sides of the square is, as shown in fig. (17.22), $\alpha\ell$ and $\beta\ell$. Thus α and β measure, as a fraction of ℓ, how far you are from the upper left corner of the square.

    Determine the magnitude and direction of the net magnetic field at the general point P. This is not difficult as long as you remain calm and keep your wits about you.

[16] While this segment does not form a closed loop and therefore does not constitute a magnetostatic current, it could be part of a more complex configuration of current that does form a closed loop — as you would see in the very next part of this problem if you had any patience. *Sheesh!*.

19. (a) Determine the magnetic field at the center of a regular polygon of k sides in which the distance from the center to each vertex is a and which carries a clockwise current I around its perimeter. See the footnote if you need a hint.[17]

    (b) Show that your result reproduces the field of a circular loop in the limit that $k \to \infty$.

    (c) Check your result for $k = 4$ against your result for #18b.

20. A finite solenoid of length ℓ, radius a, and n turns per unit length is coaxial with the z axis, extending from $z = -\ell$ to $z = 0$. A current I runs through the solenoid.

    (a) Determine the magnet field at points on the positive z axis.

    (b) There is a limit of the values of z, a, ℓ, and n against which you can easily check your answer. Figure out what this limit is and check that your result reduces to what you expect in this limit. See the footnote if you need a hint.[18]

[17] Adapt your answer to #18a to obtain a result for the contribution of each side of the polygon.

[18] You should be able to reproduce the field at the center of a ring. Just be careful about how you treat $n\ell$ as $\ell \to 0$.

17.6 Sketchy Answers

(4) $\dfrac{qB}{m}$.

(6) $\dfrac{2}{B}\sqrt{\dfrac{2mV_0}{q}}$.

(8b) Magnitude is $\dfrac{\mu_0 I\ell}{\pi(\ell^2+\frac{1}{4}a^2)}$.

(9a) The magnitudes of the forces are aIB, bIB, and cIB.

(10) Magnitude is $\dfrac{\mu_0 II'}{2\pi}\dfrac{\ell^2}{a(a+\ell)}$.

(14a) $\dfrac{\mu_0 Ir}{2\pi a^2}$ and $\dfrac{\mu_0 I}{2\pi r}\dfrac{c^2-r^2}{c^2-b^2}$ should be among your answers.

(16) $\dfrac{\mu_0 I\theta}{4\pi a}$.

(17) $\dfrac{\mu_0 I}{4\pi a}(2-\pi)$.

(18a) Magnitude is $\dfrac{\mu_0 I}{4\pi z}\dfrac{\ell}{\sqrt{\ell^2+z^2}}$.

(18b) Three symmetries give you $2^3 = 8$ terms:

$$\frac{\mu_0 I}{4\pi\ell}\Bigg(\frac{\alpha}{\beta\sqrt{\alpha^2+\beta^2}}+\frac{\alpha}{(1-\beta)\sqrt{\alpha^2+(1-\beta)^2}} + \frac{\beta}{\alpha\sqrt{\beta^2+\alpha^2}}+\frac{\beta}{(1-\alpha)\sqrt{\beta^2+(1-\alpha)^2}} + \frac{(1-\alpha)}{\beta\sqrt{(1-\alpha)^2+\beta^2}}+\frac{(1-\alpha)}{(1-\beta)\sqrt{(1-\alpha)^2+(1-\beta)^2}} + \frac{(1-\beta)}{\alpha\sqrt{(1-\beta)^2+\alpha^2}}+\frac{(1-\beta)}{(1-\alpha)\sqrt{(1-\beta)^2+(1-\alpha)^2}}\Bigg)$$

(19a) $\dfrac{\mu_0 I}{2\pi a}\,k\tan\dfrac{\pi}{k}$.

(20a) $\frac{1}{2}\mu_0 nI\left(\dfrac{z+\ell}{\sqrt{a^2+(z+\ell)^2}}-\dfrac{z}{\sqrt{a^2+z^2}}\right)$.

Chapter 18

Electrodynamics

18.1 Faraday's Law

In electrodynamics we deal with time-varying electric and magnetic fields and time-varying distributions of charge and of current. This means that in the Maxwell equations (14.1) and (14.2) we have not only to keep the terms with time derivatives of $\mathbf{E}$ and $\mathbf{B}$ but also allow that ρ and $\mathbf{j}$ may be functions of time. Because the effects of changes in the distributions of charge and current on the electric and magnetic fields propagate from their sources a finite speed (the speed of light), solving for the time-varying fields requires special techniques and the calculations are in general quite involved. Even at our level we can, however, get a lot of mileage out of the differential and integral forms of Faraday's law,

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{18.1a}$$

$$\oint_C \mathbf{E} \cdot d\mathbf{r} = -\frac{d\Phi_B}{dt} \tag{18.1b}$$

where the magnetic flux Φ_B is, by definition (14.11),

$$\Phi_B = \int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{B} \tag{18.2}$$

and is evaluated over any surface S that spans the loop C in eq. (18.1b).

While we will limit our study of electrodynamics largely to Faraday's law, what we say about Faraday's law and the effect of time-varying magnetic fields can be carried directly over to Ampère's law if we are concerned only with the effects of time-varying electric fields: if there is no current density $\mathbf{j}$ to worry about, then eqq. (14.1d) and (14.2d) reduce to

$$\nabla \times \mathbf{B} = +\epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} \qquad\qquad \oint_C \mathbf{B} \cdot d\mathbf{r} = +\epsilon_0 \mu_0 \frac{d\Phi_E}{dt}$$

which, apart from the sign difference and the nuisance factor of $\epsilon_0\mu_0$ on the right-hand side, are the "mirror image" of eqq. (18.1a) and (18.1b). So all of the results we obtain for electric fields from Faraday's law will also hold for magnetic fields and Ampère's law as long as we are careful to toss in the $\epsilon_0\mu_0$ and remember the reversal of directions due to the sign difference.

Recall now from §14.5 that

- Faraday's law tells us that a changing magnetic field induces an electric field that is a pure curl. That is, whereas the source of electrostatic fields is electric charge, so that electrostatic fields diverge from positive charge and converge into negative charge, induced electric fields are a pure circulation with no divergence.[1]

- The integral form of Faraday's law has built into it the same conventions as Stokes's theorem (2.8): when evaluating the magnetic flux Φ_B over the surface S in
$$\Phi_B = \int_S dA\, \hat{\mathbf{n}} \cdot \mathbf{B}$$
we must be sure to take the normal $\hat{\mathbf{n}}$ to the surface S to be in a right-handed sense relative to the direction in which we are integrating around the loop C in
$$\oint_C \mathbf{E} \cdot d\mathbf{r} = -\frac{d\Phi_B}{dt}$$

- The magnetic flux Φ_B is independent of the particular surface S chosen to span the loop C; we will get the same result for the magnetic flux through any surface that spans the loop. We may therefore speak simply of the magnetic flux through the loop, without reference to any particular spanning surface.

Suppose for example we have an infinitely long solenoid of radius a, within which we cause the magnetic field to decrease as time passes by decreasing the current flowing through its coil. The magnetic field, you will recall, vanishes everywhere outside the solenoid and inside is parallel to the axis of the solenoid and of uniform magnitude $B = \mu_0 nI$. At least, that was the case in magnetostatics, when the current I in the coil was constant. Since the effects of varying this current propagate from the various points on the coil only at the finite speed of light, they will reach a given location within the solenoid at differing times — the farther away a part of the coil, the longer the transit will take — , with the result that the field inside the solenoid will no longer be uniform over its cross section and would be very involved to calculate exactly.

[1] If you've forgotten what vector fields of pure divergence or pure curl look like, revisit figg. (2.14) through (2.16) on p.105.

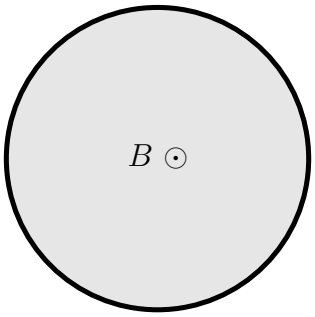


Figure 18.1: Cross Section of a Solenoid (with $+z$ $\odot$)

You may, however, have noticed that the speed of light is really fast. Really really fast. Which is very convenient: as long as the current doesn't change too quickly, and as long as the time required to travel at the speed of light from the various points on the coil to the location where we are evaluating the field is negligibly small, the effects of the changing current will reach that location effectively instantaneously, and we can therefore to a good approximation treat the magnetic field at any given instant as though it were due to the current flowing through the coil at that same instant.[2]

Let us suppose that these conditions are satisfied by our solenoid, shown in cross section in fig. (18.1) for the case of a decreasing counterclockwise current I. For the field at any given instant in the gray region interior to the solenoid we may therefore use

$$\mathbf{B} = B\,\widehat{\mathbf{z}} = \mu_0 n I\,\widehat{\mathbf{z}}$$

and likewise $B = 0$ outside the solenoid. In the cylindrical coordinates most natural to the cylindrical symmetry of the solenoid, the most general expression for the induced electric field is

$$\mathbf{E}(\mathbf{r},t) = E_r(r,\theta,z,t)\,\widehat{\mathbf{r}} + E_\theta(r,\theta,z,t)\,\widehat{\boldsymbol{\theta}} + E_\varphi(r,\theta,z,t)\,\widehat{\mathbf{z}}$$

[2]The alert reader may be disturbed by the infinite length of our solenoid: even at the speed of light, it will take a substantial time for the effects of changes in current on infinitely distant parts of the solenoid to reach the location with which we are concerned. ☹ But as long as we are making approximations, we can obviate this difficulty by supposing that the solenoid is finite and that we are well away from the field distortions at its ends. We made the solenoid infinite only in order to be able to use $B = \mu_0 n I$ in the calculations that follow.

Note also that although we are making an approximation for the sake of obtaining a result for the magnetic field, this does not mean that there is anything approximate about Faraday's law: if we were able to obtain an exact result for the magnetic field as a function of time and location, then

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

would hold exactly.

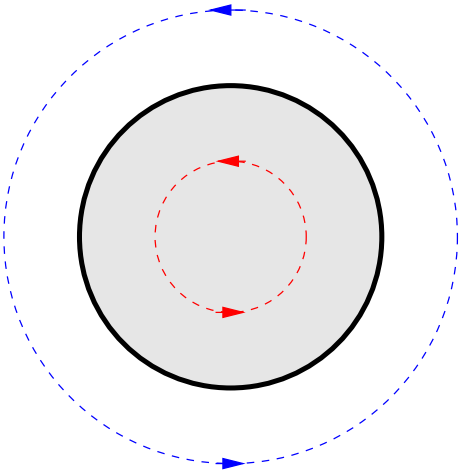


Figure 18.2: The Same Freaking Solenoid

Since this induced electric field must be a pure curl, it cannot have any radial or z component, and the usual symmetry arguments rule out its having any dependence on θ or z. Thus the induced electric field must be of the form

$$\mathbf{E}(\mathbf{r},t) = E_\theta(r,t)\,\widehat{\boldsymbol{\theta}} = E(r,t)\,\widehat{\boldsymbol{\theta}}$$

If we apply the integral form of Faraday's law to either of the two loops[3] shown in fig. (18.2), integrating counterclockwise around them, then the displacement $d\mathbf{r}$ is the element of arc $r\,d\theta\,\widehat{\boldsymbol{\theta}}$ and the line integral of the electric field works out to

$$\begin{aligned}
\oint_C \mathbf{E}\cdot d\mathbf{r} &= \oint_{\text{loop}} E(r,t)\,\widehat{\boldsymbol{\theta}}\cdot r\,d\theta\,\widehat{\boldsymbol{\theta}} \\
&= \oint_{\text{loop}} rE(r,t)\,d\theta \\
&= rE(r,t)\int_0^{2\pi} d\theta \\
&= 2\pi r E(r,t)
\end{aligned}$$

where we have noted that E does not depend on θ and can therefore be pulled out of the integral.

On the right-hand side of Faraday's law, we must now be careful to evaluate the magnetic flux using normals that are in a right-handed sense relative to the direction in which we have integrated around the loops. Since our line integrations were counterclockwise, over the circular disks that directly span

[3] Notice that we said simply *loops*, not *loops of wire*: like Gaussian surfaces or Amperian loops, the loops to which we apply Faraday's law don't have to exist physically.

the loops we must make out of the page the positive direction for magnetic flux, that is, we must take $\hat{\mathbf{n}} = +\hat{\mathbf{z}}$. For the red loop inside the solenoid we then have

$$\Phi_{B,\,red\ loop} = \int_S dA\,\hat{\mathbf{n}}\cdot\mathbf{B} = \int\limits_{\text{disk}} dA\,\hat{\mathbf{z}}\cdot B\,\hat{\mathbf{z}} = \int\limits_{\text{disk}} dA\,B = B\int\limits_{\text{disk}} dA = \pi r^2 B$$

where we have noted that the magnetic field is uniform over the cross section of the solenoid and can therefore be pulled outside the integration. For the blue loop outside the solenoid, the only difference in the flux integration is that, because B vanishes outside the solenoid, only that part of the spanning disk S that is interior to the solenoid contributes to the flux:

$$\Phi_{B,\,blue\ loop} = \pi a^2 B$$

With the above results for the line integral of the electric field and the magnetic flux, Faraday's law gives

$$\oint_C \mathbf{E}\cdot d\mathbf{r} = -\frac{d\Phi_B}{dt}$$

$$2\pi r E(r,t) = -\frac{d}{dt}\begin{cases} \pi r^2 B & (r \le a) \\ \pi a^2 B & (r \ge a) \end{cases}$$

$$= \begin{cases} -\pi r^2\,\dfrac{dB}{dt} & (r \le a) \\ -\pi a^2\,\dfrac{dB}{dt} & (r \ge a) \end{cases}$$

$$E(r,t) = \begin{cases} -\dfrac{r}{2}\dfrac{dB}{dt} & (r \le a) \\ -\dfrac{a}{2r}\dfrac{dB}{dt} & (r \ge a) \end{cases}$$

Now, the magnetic field B was in the positive z direction but decreasing, which means that dB is negative and therefore the contributions to $-d\Phi_B/dt$ on the right-hand side of the above relations is positive. Thus $E(r,t)$ on the left-hand side is positive. If we recall that $E(r,t)$ is the $\widehat{\boldsymbol{\theta}}$ component of the electric field, this means that the induced electric field is in the $+\widehat{\boldsymbol{\theta}}$ direction, that is, along the counterclockwise tangent. Thus the effect of a decreasing counterclockwise current and hence a decreasing magnetic field directed out of the page is an induced electric field that is like a counterclockwise vortex.

While we will always use the integral form of Faraday's law to determine the value of the induced electric field, it is generally easier to get the direction of the induced field from the differential form

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t}$$

In this case we have a magnetic field $\mathbf{B}$ that is out of the page but decreasing, so that $d\mathbf{B}$ is into the page. This means that the right-hand side of the above relation is out of the page. On the left-hand side, a curl that is out of the page corresponds, by the right-hand rule, to a counterclockwise induced electric field. In other words, we first figure out the vector direction of the *change* $d\mathbf{B}$ in the magnetic field, then use the right-hand rule to see which direction for the induced electric field $\mathbf{E}$ would yield a curl in the direction *opposite* to that of the change in the magnetic field.

Now that we've worked through an example, it's time to deal with your past. From a previous life in physics, you may have dealt with Faraday's law in one of the execrable forms[4]

$$V_{induced} = -\frac{\Delta\Phi_B}{\Delta t} \qquad \text{or} \qquad \mathcal{E} = -\frac{\Delta\Phi_B}{\Delta t} \tag{18.3}$$

in conjunction with a direction given by Lenz's law, which would have been stated something like

> The induced current flows in a direction that opposes the change in magnetic flux.

There are several reasons why this formulation of Faraday's law is morally reprehensible:

- The voltage (that is, the electrostatic potential) φ contributes to the electric field as a pure gradient: $\mathbf{E} = -\nabla\varphi$. Since the curl of any gradient vanishes ($\nabla \times \nabla = 0$), there is no way that a gradient can yield an electric field that has nonzero curl and consequently no way that a voltage can give rise to or account for an electric field induced according to Faraday's law. So there is in a strict sense no such thing as an "induced voltage" $V_{induced}$.

- For this reason most people instead refer to an electromotive force or EMF, usually denoted by the symbol $\mathcal{E}$. The logic behind this usage is that if Faraday's law were applied around a conducting loop like a loop of wire, the induced electric field would exert a force on the charges in the loop according to $\mathbf{F} = q\mathbf{E}$ and thus cause a current to flow around the loop. The work done on a charge q as it moved around the loop under the influence of the induced field would be

$$W_{\mathcal{E}} = \oint_{\text{loop}} \mathbf{F} \cdot d\mathbf{r} = q \oint_{\text{loop}} \mathbf{E} \cdot d\mathbf{r} = q\mathcal{E}$$

[4] If you have not had a previous course in physics and are thus untainted, you will of course not remember these things, but you should read on anyway — you will certainly also encounter induced voltages, EMF, and Lenz's law in other books, and we will perforce be referring to them ourselves in much of what follows.

If now we recall that the potential energy U associated with the true electrostatic potential φ is $U = q\varphi$ and that the change in potential energy between two points on a trajectory is always the negative of the work done ($\Delta U = -W$), we see that the work done by a true voltage has this same form:

$$W_{true\ voltage} = -\Delta U = q\Delta\varphi$$

The EMF $\mathcal{E}$ therefore not only has the same physical dimensions as voltage, but will behave just like a voltage in an electrical circuit: we can use it in Ohm's bogus law, etc., just as we would a true voltage.

But while both correct and useful within the sullied domain of engineering applications, the concept of EMF falls far short of reflecting the full physical import and much more general application of Faraday's law. In pure terms, Faraday's law tells us that a time-varying magnetic field induces an electric field. Period. There is no need for a conducting loop or even for any charge to be present; Faraday's law applies even in a vacuum. And what is induced is not a voltage, not a current, not a force; what is induced is an electric field.

- Both the magnitude and the direction of the induction effects are given by the differential form

  $$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

  of Faraday's law or, equivalently, its integral form

  $$\oint_C \mathbf{E} \cdot d\mathbf{r} = -\frac{d\Phi_B}{dt}$$

  There is no need to give a separate prescription like Lenz's law for the direction, nor is there anything in Lenz's law that is not already in Faraday's law (and far more succinctly and accurately expressed by it, at that).

  Moreover, as just noted, Faraday's law applies far more generally than to the induced currents and physical loops of wire conjured up by Lenz's law.

That said, certain people — and we all know who *they* are — expect you to analyze induction effects according to Lenz's law and the adulterated version (18.3) of Faraday's law. If, for example, we were to apply Lenz's law to our analysis of the varying solenoid field, the reasoning would be as follows: The magnetic field is out of the page but decreasing, so the change in flux is into the page. To oppose this change, a current that would create a

magnetic field out of the page would therefore be induced. By the right-hand rule, this requires a counterclockwise induced current. This is where Lenz's law stops. But the direction of the induced electric field will always be the same as that of the induced current, since by $\mathbf{F} = q\mathbf{E}$ positive charge and thus the current would move in the same direction as the electric field.[5]

If you have to calculate an EMF, you simply use eq. (18.3),

$$\mathcal{E} = -\frac{d\Phi_B}{dt}$$

or

$$\mathcal{E} = \oint d\mathbf{r} \cdot \mathbf{E}$$

which by Faraday's law are of course equivalent. In our solenoid example, integrating counterclockwise around the red loop would give

$$\begin{aligned}
\oint_{\text{loop}} d\mathbf{r} \cdot \mathbf{E} &= \oint_{\text{loop}} r\, d\theta\, \widehat{\boldsymbol{\theta}} \cdot E(r,t)\, \widehat{\boldsymbol{\theta}} \\
&= \oint_{\text{loop}} r\, d\theta\, \widehat{\boldsymbol{\theta}} \cdot \left(-\frac{r}{2}\, \frac{dB}{dt}\, \widehat{\boldsymbol{\theta}} \right) \\
&= -\frac{r^2}{2}\, \frac{dB}{dt} \int_0^{2\pi} d\theta \\
&= -\pi r^2\, \frac{dB}{dt}
\end{aligned}$$

Alternatively, we could use the adulterated version (18.3) of Faraday's law: since the flux through the loop was, as we saw, $\Phi_B = \pi r^2 B$, the EMF would be

$$\mathcal{E} = -\frac{d\Phi_B}{dt} = -\frac{d}{dt}(\pi r^2 B) = -\pi r^2\, \frac{dB}{dt}$$

In whichever way obtained, the magnitude of the EMF around the loop is

$$\mathcal{E} = \pi r^2 \left| \frac{dB}{dt} \right|$$

and it would, as we found both by Faraday's law proper and by Lenz's law, induce a counterclockwise current. In a circuit, this EMF would therefore be equivalent to a battery of voltage

$$\pi r^2 \left| \frac{dB}{dt} \right|$$

facing in the direction shown in fig. (18.3).

[5] When we reasoned out our directions from Faraday's law proper, the order of these last two steps in the reasoning was reversed: we directly concluded that the direction of the induced electric field was counterclockwise. If there were a conducting loop present, there would therefore be a counterclockwise current induced around this loop. But, again, the loop does not have to exist physically, let alone be conducting.

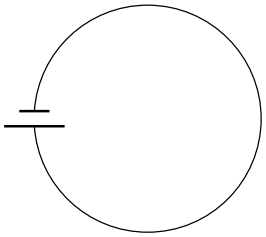


Figure 18.3: The Battery Mimicking the EMF

Finally, be mindful that if a conducting loop is present, having several turns (that is, windings) in the loop effectively increases its area and therefore the flux through it: a loop of N turns will have N times as much flux through it as a simple, single loop of the same cross-sectional area.

18.2 Ye Olde Sliding Bar

Since the sort of people who ask you about these sorts of things expect you to do your analysis by Lenz's law and the adulterated version of Faraday's law, that is how we will do our analysis.

Suppose you have a U-shaped conducting track, along which a conducting bar is sliding, as shown in fig. (18.4). In this case, the change in flux is due to the change in the area of the loop through which the flux is passing. If the bar is moving at velocity v, the track has width ℓ, and the magnetic field B throughout the region is constant and out of the page, then the increase in the area of the loop formed by the track and bar in time interval dt is $dA = \ell\,(v\,dt)$ and the corresponding increase in flux coming out of the page is thus $d\Phi_B = B\,dA = B\ell v\,dt$. The EMF induced around the loop is therefore

$$\mathcal{E} = -\frac{d\Phi_B}{dt} = -\frac{B\ell v\,dt}{dt} = -B\ell v$$

Since we are acquiring additional flux out of the page, the change in flux

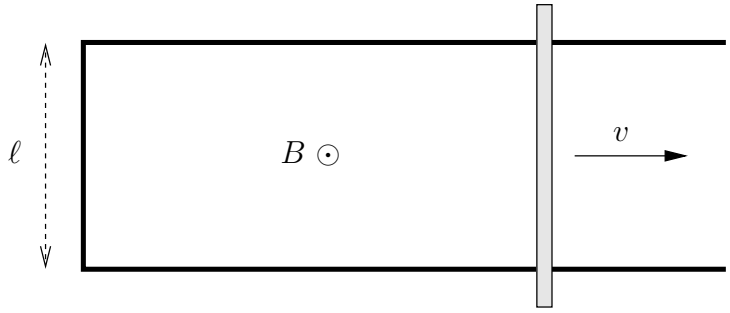


Figure 18.4: The Old Sliding Bar on the U-Shaped Track

is out of the page. To oppose this change, a current that would create a magnetic field into the page would therefore be induced. By the right-hand rule, this requires a clockwise induced current.

There are many games that can be played with bars sliding on tracks; the analysis of all of them is very similar.

Because Faraday's law does not require the presence of a physical loop, it is also possible to apply it to the motion of an isolated conducting bar without a track. The analysis is essentially the same as for the bar on the track: we can apply Faraday's law to an imaginary loop, around which the EMF will again be of magnitude $B\ell v$, so that the effective voltage difference between the ends of the bar will be $B\ell v$. If this analysis seems a bit hokey, there is another way to look at this situation: as the bar moves through the magnetic field, the charges on it will experience a magnetic force, with the magnetic forces on the positive and negative charges being in opposite directions. For the bar and magnetic field shown in fig. (18.4) (without the track, of course), positive charges will, by $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$, experience a magnetic force down the page ($\downarrow$), and negative charges a magnetic force up the page ($\uparrow$). The magnetic force will thus tend to separate the charge on the bar. This separation will, however, result in an electrical attraction between the opposite charges that will tend to pull them back together. Equilibrium will be established when the magnitudes of the electric and magnetic forces are balanced. If charge $+q$ has accumulated at the bottom of the bar and charge $-q$ at its top, the magnitude of the electrical attraction will be

$$F_{electric} = \frac{1}{4\pi\epsilon_0}\frac{q^2}{\ell^2}$$

while the magnetic force on each charge will be of magnitude

$$F_{magnetic} = qvB\sin\frac{\pi}{2} = qvB$$

Also, the electrical potential energy of the pair of charges will be

$$U = \frac{1}{4\pi\epsilon_0}\frac{q^2}{\ell}$$

If, using $U = q\varphi$, we regard this potential energy as the product of one charge with the effective voltage due to the other charge, then we obtain for the effective voltage difference between the charges

$$V_{\mathit{eff}} = \frac{1}{4\pi\epsilon_0}\frac{q}{\ell}$$

Equating the electric and magnetic forces and solving for this effective voltage, we arrive at the same result we obtained by Faraday's law:

$$\frac{1}{4\pi\epsilon_0}\frac{q^2}{\ell^2} = qvB$$

$$\left(\frac{1}{4\pi\epsilon_0}\frac{q}{\ell}\right)\frac{q}{\ell} = qvB$$
$$V_{\text{eff}}\,\frac{q}{\ell} = qvB$$
$$V_{\text{eff}} = B\ell v$$

18.3 Generators & Motors

Consider a rectangular loop of wire rotating in a region of constant magnetic field, as shown in fig. (18.5): the loop is of area A and is rotating at a constant angular rate ω such that at the moment shown in the figure the left side is coming out of the page and the right side is going into the page. The constant magnetic field B in the region is out of the page. Since we will need some way to "harvest" the induced EMF, we have made a small break in the loop and connected to it the pair of blue wires shown in the figure.

In our integration (18.2) for the magnetic flux, the magnetic field and the spanning rectangular surface can both be treated as constant, with the varying quantity being the angle between the magnetic field and the normal $\hat{\mathbf{n}}$ to the rectangle. If we take out of the page as the initial direction for $\hat{\mathbf{n}}$, then this angle starts at zero at the moment shown in fig. (18.5) and will be given more generally by ωt. The EMF induced around the loop is then

$$\begin{aligned}\mathcal{E} &= -\frac{d\Phi_B}{dt}\\ &= -\frac{d}{dt}\int_{\text{loop}} dA\,\hat{\mathbf{n}}\cdot\mathbf{B}\\ &= -\frac{d}{dt}\int_{\text{loop}} dA\,B\cos\omega t\end{aligned}$$

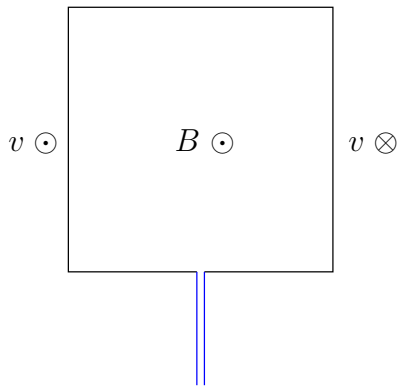


Figure 18.5: A Rudimentary Generator

$$\begin{aligned}
&= -\frac{d}{dt}\left(B\cos\omega t \int\limits_{\text{loop}} dA \right) \\
&= -\frac{d}{dt}(BA\cos\omega t) \\
&= \omega BA\sin\omega t
\end{aligned}$$

Thus we get a sinusoidally varying EMF — AC electricity. All you have to do to generate such electricity is rotate a loop of wire in a magnetic field. Real generators are, of course, far more sophisticated than our rectangular loop, but it is by this simple principle that they all operate.

If the generator is connected across an effective resistance R, by Ohm's bogus law the corresponding current will be

$$I = \frac{V}{R} = \frac{\omega BA}{R}\sin\omega t \tag{18.4}$$

and the electrical power supplied by the generator will be

$$P_{generator} = \frac{V^2}{R} = \frac{(\omega BA)^2}{R}\sin^2\omega t \tag{18.5}$$

This power is not, of course, for free: work must be done to keep the loop rotating. At the moment shown in fig. (18.5), the magnetic flux is out of the page and will, as the loop continues to rotate, decrease, so that the change in the flux will be into the page. To oppose this change by making a contribution to the magnetic flux directed out of the page, the induced current will flow counterclockwise. This loop of current constitutes a magnetic dipole, the dipole moment $\boldsymbol{\mu}$ of which is, by eqq. (17.25) and (18.4),

$$\boldsymbol{\mu} = IA\,\widehat{\mathbf{n}} = \left(\frac{\omega BA}{R}\sin\omega t\right) A\,\widehat{\mathbf{n}} = \frac{\omega BA^2}{R}\sin\omega t\,\widehat{\mathbf{n}} \tag{18.6}$$

where this $\widehat{\mathbf{n}}$ is (see fig. (17.9) on p.911) the same as the normal $\widehat{\mathbf{n}}$ to the loop. By eqq. (17.24) and (18.6), the magnitude of the magnetic torque on this dipole will be

$$\begin{aligned}
\tau &= |\boldsymbol{\mu}\times\mathbf{B}| \\
&= \mu B\sin\angle(\widehat{\mathbf{n}},\mathbf{B}) \\
&= \mu B\sin\omega t \\
&= \left(\frac{\omega BA^2}{R}\sin\omega t\right) B\sin\omega t \\
&= \frac{\omega B^2A^2}{R}\sin^2\omega t
\end{aligned}$$

Although at the instant $t = 0$ shown in fig. (18.5) this torque vanishes, you can see that at a slightly later time, when the normal $\hat{\mathbf{n}}$ will have tilted slightly the right, the direction of $\hat{\mathbf{n}} \times \mathbf{B}$ and hence of the magnetic torque will be down the page. This means that the magnetic torque will be against the rotation of the loop, so that you will have to supply an equal torque in the opposite direction in order to keep the loop rotating. By eq. (7.26), the power associated with the torque you apply is

$$P_{you} = \tau\omega = \left(\frac{\omega B^2 A^2}{R}\sin^2\omega t\right)\omega = \frac{(\omega BA)^2}{R}\sin^2\omega t$$

Thus to keep the loop rotating you have to do work at a rate exactly equal to the electrical power output of the generator.

An electric motor is just an electric generator run, so to speak, backward: instead of supplying the torque needed to keep the loop rotating and induce the sinusoidal EMF and current that constitute AC electricity, you put a current through the loop by connecting it to a voltage source like an electrical outlet. The consequent magnetic torque on the loop causes it to rotate, and that torque and the associated power can be used to do the mechanical work of turning gears or wheels or whatever.

18.4 On the Issue of Time Derivatives

Now that you have had some experience working with Faraday's law, let us revisit the issue of the total time derivative in its integral form

$$\oint_C \mathbf{E}\cdot d\mathbf{r} = -\frac{d\Phi_b}{dt} \tag{18.7}$$

versus the partial time derivative in its differential form

$$\nabla\times\mathbf{E} = -\frac{\partial\mathbf{B}}{\partial t} \tag{18.8}$$

In general, there are three kinds of contributions to the change $d\Phi_B$ in the magnetic flux Φ_B during a time interval dt:

- Directly from the change $d\mathbf{B}$ in the magnetic field as a function of time.
- From the change dC in the perimeter C of the surface S through which the flux is being calculated, due to translation of entire perimeter or to changes in its shape or extent.

  These sorts of changes would be generated when the entire surface moves rigidly through space, or by a rubbery perimeter that expands or contracts or changes shape.

- From the change dS in the surface spanning the perimeter C due to "flexing" of S.

  These sorts of changes would be like those in the cheese on an open-face grilled cheese sandwich in the broiler: there is no change in the perimeter of the cheese, but as it cooks its surface bubbles outward in some places and caves inward in others.

Because we are looking at the infinitesimal change $d\Phi_B$ in flux during the infinitesimal time interval dt, we have only to add the individual contributions from each of these three sources of change, keeping the other two of $\{\mathbf{B},\, C,\, S\}$ constant: changes in Φ_B due to simultaneous changes in two or more of these three animals will be second- or third-order small. Thus

$$\begin{aligned}\text{Total change in } \Phi_B = \quad & \text{Change due to } d\mathbf{B} \text{ with } C \text{ and } S \text{ constant} \\ & + \text{Change due to } dC \text{ with } \mathbf{B} \text{ and } S \text{ constant} \\ & + \text{Change due to } dS \text{ with } \mathbf{B} \text{ and } C \text{ constant}\end{aligned}$$

which we will express as

$$\begin{aligned} d\Phi_B &= d\Big(\int_S dA\,\hat{\mathbf{n}}\cdot\mathbf{B}\Big) \\ &= \int_S dA\,\hat{\mathbf{n}}\cdot d\mathbf{B} + \int_{dC} dA\,\hat{\mathbf{n}}\cdot\mathbf{B} + \int_{dS} dA\,\hat{\mathbf{n}}\cdot\mathbf{B}\end{aligned}$$

The contributions to the rate of change of $d\Phi_B$ are thus

$$\frac{d\Phi_B}{dt} = \frac{1}{dt}\int_S dA\,\hat{\mathbf{n}}\cdot d\mathbf{B} + \frac{1}{dt}\int_{dC} dA\,\hat{\mathbf{n}}\cdot\mathbf{B} + \frac{1}{dt}\int_{dS} dA\,\hat{\mathbf{n}}\cdot\mathbf{B} \tag{18.9}$$

Being entirely due to the explicit time variation of the magnetic field, the first of these three contributions is that of $\partial\mathbf{B}/\partial t$:

$$\int_S dA\,\hat{\mathbf{n}}\cdot\frac{\partial\mathbf{B}}{\partial t} \tag{18.10}$$

This is the expected contribution to the right-hand side of Faraday's law.

Now consider the contribution to $d\Phi_b$ from dS. Since S and $S+dS$ share the same perimeter C, together they will form a closed surface.[6] But by $\nabla\cdot\mathbf{B} = 0$ the net outward flux through any closed surface vanishes. The flux out through $S+dS$ must therefore equal the flux in through S, so that

$$0 = \int_{S+dS} dA\,\hat{\mathbf{n}}\cdot\mathbf{B} - \int_S dA\,\hat{\mathbf{n}}\cdot\mathbf{B}$$

[6] Or, if you will, a set of closed surfaces, since S may bubble outward in some places and cave inward in others. But the argument we are making can then be applied to each of these closed surfaces individually.

$$= \int_{dS} dA\, \hat{\mathbf{n}} \cdot \mathbf{B} \tag{18.11}$$

In eq. (18.9) there is therefore no contribution to $d\Phi_b/dt$ from dS.

Working out the contribution of dC to $d\Phi_b/dt$ is more involved. The change dC in the perimeter of S will constitute a ribbon of infinitesimal width. Recall now from problem #20 of Chapter 1 (on p.81) that the area of a parallelogram of edges $\mathbf{a}$ and $\mathbf{b}$ is given by $|\mathbf{a} \times \mathbf{b}|$. The area elements dA of our ribbon dC may therefore be expressed as $|d\boldsymbol{\ell} \times d\mathbf{r}|$, where $d\boldsymbol{\ell}$ is an infinitesimal segment of the perimeter and $d\mathbf{r} = \mathbf{v}\, dt$ is the displacement of this segment during the time interval dt due to the motion or flexing of the perimeter C. Since $d\boldsymbol{\ell} \times d\mathbf{r}$ is also, very conveniently, perpendicular to the plane of dA, we have

$$dA\, \hat{\mathbf{n}} = d\boldsymbol{\ell} \times d\mathbf{r} = d\boldsymbol{\ell} \times \mathbf{v}\, dt$$

so that we can rewrite the contribution of dC to $d\Phi_B/dt$ as

$$\begin{aligned} \frac{1}{dt} \int_{dC} dA\, \hat{\mathbf{n}} \cdot \mathbf{B} &= \frac{1}{dt} \int_{dC} d\boldsymbol{\ell} \times \mathbf{v} \cdot \mathbf{B} \\ &= \oint_C d\boldsymbol{\ell} \times \mathbf{v} \cdot \mathbf{B} \\ &= \oint_C \mathbf{v} \times \mathbf{B} \cdot d\boldsymbol{\ell} \end{aligned} \tag{18.12}$$

where we have noted that once the dt is divided out, our integration over the ribbon has collapsed down to an integration over the perimeter, and that, as you can straightforwardly (if rather tediously) verify by just doing out the components,

$$d\boldsymbol{\ell} \times \mathbf{v} \cdot \mathbf{B} = \mathbf{v} \times \mathbf{B} \cdot d\boldsymbol{\ell}$$

Putting together eqq. (18.10) through (18.12), we arrive at

$$\frac{d\Phi_B}{dt} = \int_S dA\, \hat{\mathbf{n}} \cdot \frac{\partial \mathbf{B}}{\partial t} + \oint_C \mathbf{v} \times \mathbf{B} \cdot d\boldsymbol{\ell}$$

which, when used in the integral form (18.7) of Faraday's law, gives

$$\oint_C \mathbf{E} \cdot d\mathbf{r} = -\frac{d\Phi_b}{dt} = -\int_S dA\, \hat{\mathbf{n}} \cdot \frac{\partial \mathbf{B}}{\partial t} - \oint_C \mathbf{v} \times \mathbf{B} \cdot d\boldsymbol{\ell}$$

Since $d\mathbf{r}$ and $d\boldsymbol{\ell}$ both represent segments of the perimeter C, we can consolidate our notation by using $d\mathbf{r}$ for both. Combining the two line integrals then yields

$$\oint_C (\mathbf{E} + \mathbf{v} \times \mathbf{B}) \cdot d\mathbf{r} = -\int_S dA\, \hat{\mathbf{n}} \cdot \frac{\partial \mathbf{B}}{\partial t}$$

If we use Stokes's theorem

$$\oint_C \mathbf{X} \cdot d\mathbf{r} = \int_S dA\, \hat{\mathbf{n}} \cdot \nabla \times \mathbf{X}$$

to rewrite the left-hand side, this becomes

$$\int_S dA\, \hat{\mathbf{n}} \cdot \nabla \times (\mathbf{E} + \mathbf{v} \times \mathbf{B}) = -\int_S dA\, \hat{\mathbf{n}} \cdot \frac{\partial \mathbf{B}}{\partial t}$$

Since this relation must hold for all surfaces S, we conclude that

$$\nabla \times (\mathbf{E} + \mathbf{v} \times \mathbf{B}) = -\frac{\partial \mathbf{B}}{\partial t} \tag{18.13}$$

This is certainly not the differential form (18.8) of Faraday's law. But it is what we would expect physically if, as we noted in footnote 11 on p.757, we want the total time derivative $d\Phi_B/dt$ to take into account the effect of motion or morphing of the surface S: while changes dS in the surface at points other than along the perimeter have no effect, when there is motion of the perimeter of S a charge q on it would experience not only the electric force $q\mathbf{E}$ but, because of the velocity $\mathbf{v}$ of the perimeter at q's location, a magnetic force $q\mathbf{v} \times \mathbf{B}$. And changing $\mathbf{F} = q\mathbf{E}$ to

$$q\mathbf{E} + q\mathbf{v} \times \mathbf{B} = q(\mathbf{E} + \mathbf{v} \times \mathbf{B})$$

is equivalent to changing $\mathbf{E}$ to $\mathbf{E} + \mathbf{v} \times \mathbf{B}$ — exactly as accounted for on the left-hand side of eq. (18.13).

Similar arguments may be applied to the integral and differential forms of Ampère's law,

$$\nabla \times \mathbf{B} = \mu_0\, \mathbf{j} + \epsilon_0 \mu_0 \frac{\partial \mathbf{E}}{\partial t} \qquad \oint_C \mathbf{B} \cdot d\mathbf{r} = \mu_0\, I_{enclosed} + \epsilon_0 \mu_0 \frac{d\Phi_E}{dt}$$

with the differences that the roles of $\mathbf{E}$ and $\mathbf{B}$ are interchanged, that we have a factor of $\epsilon_0\mu_0$ with the flux term, and that, because the flux term on the right-hand side now occurs with a positive sign, the contribution from dP would have the opposite sign. So instead of

$$\mathbf{E} \to \mathbf{E} + \mathbf{v} \times \mathbf{B} \tag{18.14a}$$

we would have

$$\mathbf{B} \to \mathbf{B} - \epsilon_0 \mu_0 \mathbf{v} \times \mathbf{E} \tag{18.14b}$$

This is not as easy to make sense of as $\mathbf{E} + \mathbf{v} \times \mathbf{B}$, and in fact it turns out that the proper explanation of both (18.14a) in Faraday's law and (18.14b) in Ampère's law is the Lorentz covariance of electromagnetism: relativity is built into the Maxwell equations, and while unfortunately we are not in a

position to work through the calculation, it turns out that under the Lorentz transform, just as spatial distances and time intervals are mixed together as you go from one reference frame to another, so that what was purely a spatial distance in one frame is partly a time interval in another and vice versa, what is an electric field in one frame will be partly a magnetic field in another and vice versa — exactly as given by eqq. (18.14).

18.5 Problems

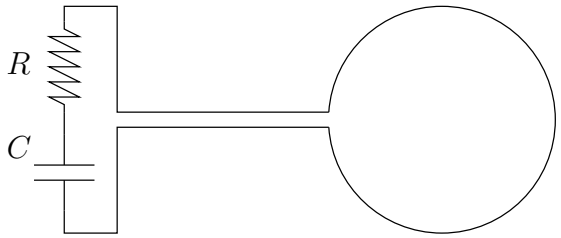


Figure 18.6: Problem 1

1. Fig. (18.6) is a somewhat oversimplified "schematic" of a device that can be used to measure magnetic field strength: a loop of wire of area A (actually, a coil of N turns, each turn of area A) is connected to a resistance R and capacitance C. With the capacitor initially uncharged, the loop is moved from a region where there is little or no magnetic field into the region of the field to be measured, with the coil oriented as perpendicularly to the field as possible. The induced EMF and current lead to an accumulation of charge and voltage across the plates of the capacitor, and it is this voltage that the device directly measures.[7]

   If the overall voltage accumulated on the capacitor is V_c, what is the magnitude of the magnetic field?

[7] We have omitted from fig. (18.6) a required circuit element known as a diode, a semiconductor device that allows current to flow in only one direction. Without the diode, the capacitor would simply discharge backward through the loop as the induced EMF died away.

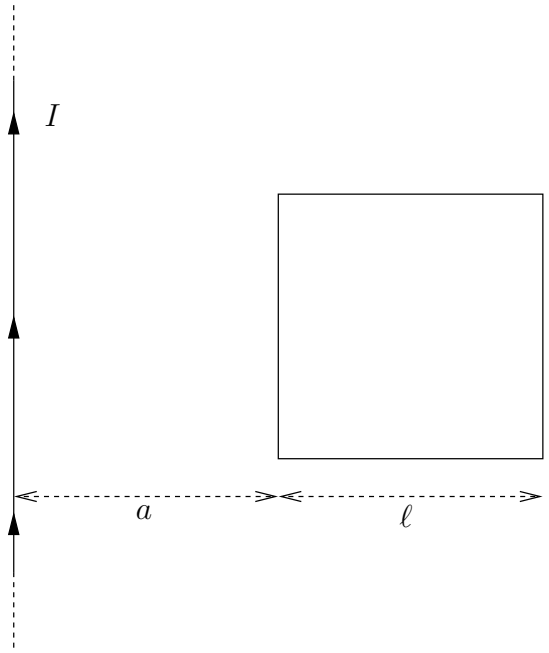


Figure 18.7: Problem 2

2. Fig. (18.7) shows an infinite straight wire carrying a current I up the page (that is, up the page is the positive direction for the current flow). A perpendicular distance a from the wire, and in a plane with it, is a conducting square of side ℓ, as shown in fig. (18.7). The current I is given as a function of the time t by

$$I = I_0\left(\frac{t}{\tau}\right)^2$$

where I_0 and τ are positive constants.

   (a) Determine the EMF induced around the square as a function of time.

   (b) What assumptions did you make about the values of I_0, τ, ℓ, and a? See the footnote if you need a hint.[8]

   (c) Determine the direction of the current induced around the square as a function of time. Be sure to consider both $t > 0$ and $t < 0$.

[8]Remember that changes in the magnetic field propagate from their sources at the speed of light.

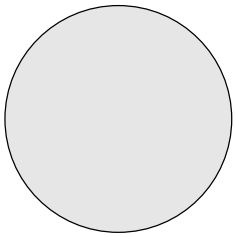


Figure 18.8: Problem 3

3. Fig. (18.8) shows a cross section of a long, thin solenoid of radius a, length ℓ, and n turns per unit length. Why we even bothered with this figure is anybody's guess. We suppose you could use it to draw a smiley with a mustache on it. Anyway, the current flowing around the solenoid is changing as time passes, so that the magnetic field B inside the solenoid is also varying.

   (a) If the current $I(t)$ is treated as a given, determine the induced electric field at a distance r from the center of the solenoid. Be sure to consider both $r < a$ and $r > a$.

   (b) Determine the direction of the induced electric field if the current in the solenoid is counterclockwise and decreasing.

   (c) In deriving your result for the magnitude of the induced electric field, you doubtless assumed that the changes in the magnetic field were instantaneous, when in reality all changes in electromagnetic fields propagate from their sources — in this case the wires of the solenoid — only at the finite speed of light. What must be true of $I(t)$, a, n, ℓ, and r in order for your results for the magnitude of the induced electric field to be a good approximation?

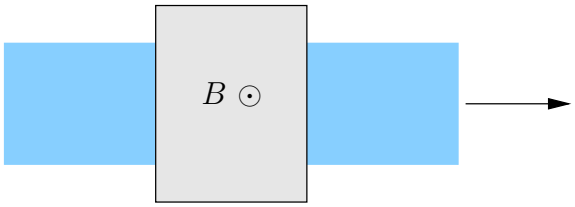


Figure 18.9: Problem 4

4. Fig. (18.9) is a rather lame representation of a blue metal plate be drawn to the right, past a gray magnet that gives rise to a field that is approximately constant and out of the page in the gray region and zero elsewhere. The changes in flux through the various regions of the metal plate give rise to induced voltages, and, since the plate is a conductor, these induced voltages in turn give rise to induced currents. These swirling currents are known as eddy currents. Determine the direction, clockwise or counterclockwise, of the eddy currents in the regions of the plate that are entering and leaving the shaded region of magnetic field.

   (This effect is the basis of magnetic braking or magnetic damping. You may have seen balances that use magnetic damping to damp out oscillations. And on some stationary bikes the pedaling resistance is provided magnetically: Ultimately you are rotating a metal disk surrounded by lots of magnets. The resistance is increased by moving the magnets closer to the disk, thereby increasing the strength of the flux and the magnitude of the induced current. Although the disk is a conductor, it has some resistance, and the power you are putting into pedaling is being converted into heat in the disk according to $I^2_{induced}R$.)

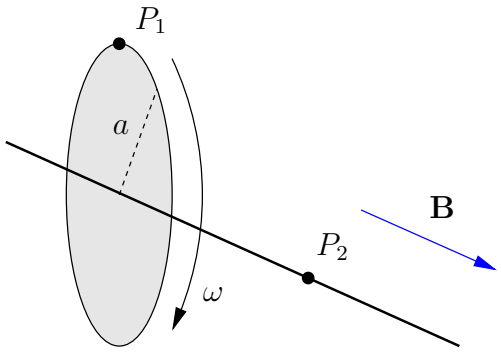


Figure 18.10: Problem 5

5. Believe it or not, fig. (18.10) shows a conducting disk of radius a rotating around a conducting axis at angular velocity ω. This simple arrangement constitutes what is known as a homopolar generator: the disk is rotated in a region where there is a constant magnetic field parallel to the axis, and brushes with metal bristles, in contact with the points P_1 and P_2, tap the resulting EMF.

   (a) Show that the induced EMF between points P_1 and P_2 is

   $$\mathcal{E} = \tfrac{1}{2}\omega B a^2$$

   See the footnote if you need a hint.[9]

   (b) If the magnetic field is in the direction indicated by the blue arrow and the rotation of the disk is in a right-handed sense relative to this direction, which point is at the higher voltage, P_1 or P_2?

[9] Think in terms of the area traced out as the disk makes one full revolution.

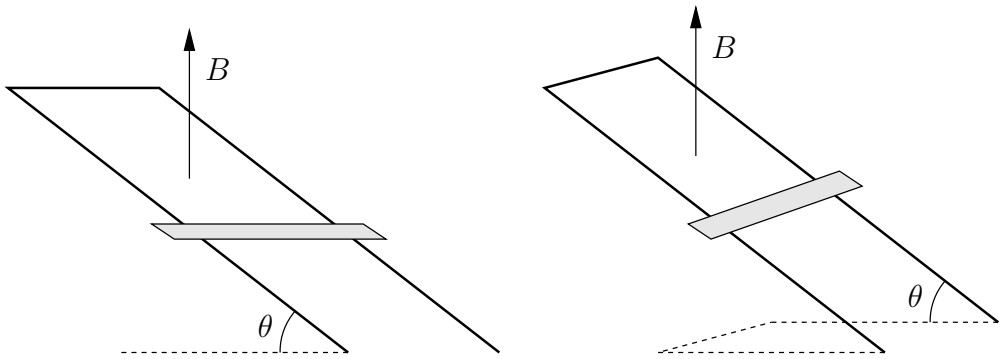


Figure 18.11: Problem 6

6. The left side of fig. (18.11) shows a conducting bar of mass m and length ℓ sliding down a conducting track inclined at angle θ to the horizontal from a bizarrely medieval perspective. The right side of the figure shows the very same bar and track from a more Renaissance perspective. The track has no resistance, and the resistance of the bar is R. Everywhere in the region of the track the magnetic field has constant magnitude B and is vertically upward.

   (a) Show that the magnetic force on the bar when it is moving at speed v is
   $$F_{mag} = \frac{B^2\ell^2 v\cos\theta}{R}$$

   (b) When asked to determine the speed of the bar after it has traveled, starting from rest, a distance x down the track, someone reasons as follows: Since the net force on the bar is
   $$F_{net} = mg\sin\theta - \frac{B^2\ell^2 v\cos\theta}{R}$$
   the net work done on the bar as it moves a distance x down the track is
   $$W = \int_0^x dx\,(mg\sin\theta - \frac{B^2\ell^2 v\cos\theta}{R}) = (mg\sin\theta - \frac{B^2\ell^2 v\cos\theta}{R})\,x$$
   And since the bar starts from rest,
   $$W = \Delta K = \tfrac{1}{2}mv^2 - \tfrac{1}{2}mv_0^2 = \tfrac{1}{2}mv^2$$
   gives
   $$\tfrac{1}{2}mv^2 = (mg\sin\theta - \frac{B^2\ell^2 v\cos\theta}{R})\,x$$
   and hence, when we solve for v,
   $$v = -\frac{B^2\ell^2 x\cos\theta}{mR} + \sqrt{\left(\frac{B^2\ell^2 x\cos\theta}{mR}\right)^2 + 2gx\sin\theta}$$
   Explain why this reasoning is moronic.

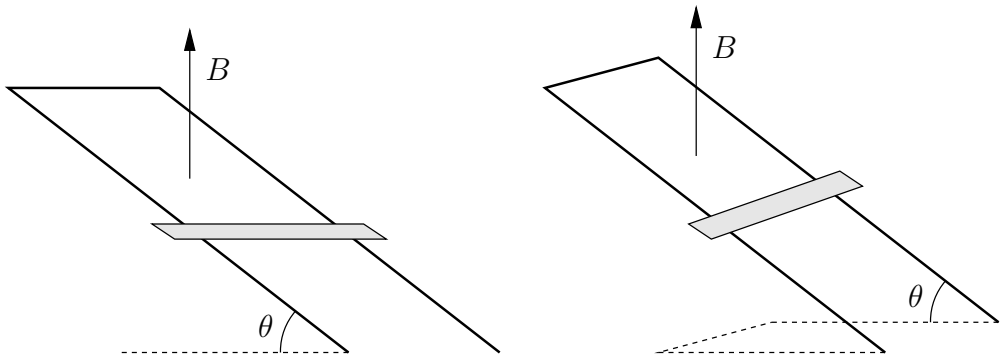


Figure 18.11: Problem 6

(c) Show that the terminal velocity attained by the rod is

$$v = \frac{mgR\tan\theta}{B^2\ell^2}$$

(d) Show that the velocity of the rod as a function of time is given by

$$v = \frac{mgR\tan\theta}{B^2\ell^2}\left(1 - \exp\left(-\frac{B^2\ell^2 v\cos\theta}{mR}\,t\right)\right)$$

(e) Describe the motion.

(f) Make physical sense of the dependence of the terminal velocity on m, R, θ, B, and ℓ.

(g) Make physical sense of the velocity in the limit $R \to 0$, that is, as the bar becomes a superconductor.

(You should now be able to explain superconducting magnetic levitation.)

(h) How would your results and answers to the various parts of this problem change if the track had a cross piece at the bottom as well as at the top?

7. Consider a sliding bar just like that of §18.2, but with a special device built into the track that keeps a constant current I flowing around the circuit and in particular through the bar. The track has no resistance; the resistance of the bar is R.

   (a) Determine the velocity of the bar as a function of time, taking the bar to start from rest at time $t = 0$.

   (b) Determine, as a function of the time t, the voltage that this device must supply in order to keep the current constant.

   (c) We have maintained all along that magnetic forces never do any work, so what the $****$ is going on here? What is the source of the bar's kinetic energy?

8. Now consider a sliding bar just like that of §18.2, but with a special device built into the track that applies a constant voltage V_b around the circuit — a device known as ... *a battery!*[10] The track has no resistance; the resistance of the bar is R.

   (a) Show that if the bar starts from rest at time $t = 0$ its velocity as a function of time is given by

   $$v = \frac{V_b}{B\ell}\left(1 - \exp\left(-\frac{B\ell^2}{mR}t\right)\right)$$

   See the footnote if you need a hint.[11]

   (b) Describe the motion.

   (c) Make physical sense of the dependence of this velocity on B, m, ℓ, R, and V_b.

   (d) Show that power is conserved in this circuit.

9. (The partner to the magnetic fringe-field problem #15 of Chapter 17.) Recall that the field of an idealized parallel-plate capacitor is taken to go abruptly from σ/ϵ_0 in between the plates to zero outside the plates. Use Faraday's law to show that this cannot in fact happen.

[10] This should be read with the same climactic emphasis as "Bring us ... *a shrubbery!*"

[11] Remember that the force on the bar depends on the current and that the current in turn depends on the *net* voltage around the circuit.

18.6 Sketchy Answers

(1) $\dfrac{RCV_c}{NA}$.

(2a) $\dfrac{\mu_0 I_0}{\pi}\,\ell \ln\dfrac{\ell}{a}\,\dfrac{t}{\tau^2}$.

(7a) $\dfrac{I\ell B}{m}t$.

(7b) $IR\left(1+\dfrac{\ell^2 B^2}{mR}t\right)$.

Chapter 19

More DC Circuits

19.1 Inductance

Running a current through a loop of wire will create a magnetic field and thus a magnetic flux through the loop. Changing this current will therefore induce an electric field and EMF around the loop. This inductive effect of a circuit element on itself is called self-inductance. If instead we have two loops of wire, changing the current flowing through one loop will change the flux through the other loop and thus induce an electric field and voltage around that other loop. The current and magnetic field so induced in that other loop will, by a sort of feedback, give rise to a change in flux through the first loop, and so on. This inductive effect of the two loops on each other is called mutual inductance. Our study of inductance will be restricted to self-inductance.

An inductor is a circuit element that can give rise to a time-varying magnetic flux and hence to induction effects. An inductor is usually a more or less lengthy coil of many turns of wire: changing the current through the coil changes the magnetic field and hence the magnetic flux through the coil, and this changing magnetic flux induces a voltage across the coil. In principle, since any current will give rise to a magnetic field and thus an induction effect, all circuit elements have an inductance. In practice, however, these inductance effects are negligibly small except for circuit elements specifically designed for the purpose.

The magnetic field B within an inductor, and hence the magnetic flux Φ_B through it, will be proportional to the current I through it. The constant of proportionality is defined to be the inductance L:

$$\Phi_B = LI \tag{19.1}$$

If we use eq. (19.1) in the adulterated version

$$\mathcal{E} = -\frac{d\Phi_B}{dt}$$

of Faraday's law, the EMF induced across the inductor may be expressed as

$$\mathcal{E} = -L\,\frac{dI}{dt} \tag{19.2}$$

The negative sign indicates that a time-varying current will back up a voltage $L\dot{I}$ across it:[1] phrased anthropomorphically, inductors don't like change and will resist it by limiting the rate at which the current changes.

To understand the sign on the EMF, think of a resistor: when we write $V = IR$, we mean that the charge passing through the resistor experiences a *drop* IR in voltage, corresponding to its losing energy inside the resistor. That's why in the loop rule we use $-IR$ for the voltage change when we cross a resistor with the current. Eq. (19.2) differs from Ohm's bogus law in that the negative is explicit, but the effect is in the same direction: there is a drop $L\dot{I}$ in voltage across the inductor. It is just that the charge passing through the inductor, instead of losing energy to heat, is, as we will see in the following paragraphs, going "uphill" in a kind of magnetic potential energy.

Within an infinitely long solenoid of n turns per unit length, the magnetic field is $B = \mu_0 nI$, and this will to a good approximation also hold for a finite solenoid of cross-sectional area A and length ℓ as long as it is long and narrow enough. Since the total number of turns in such a finite solenoid is $N = n\ell$, the total flux through it is

$$\Phi_B = NBA = (n\ell)(\mu_0 nI)A = \mu_0 n^2 A\ell I$$

Comparing this with $\Phi_B = LI$ yields

$$L = \mu_0 n^2 A\ell \tag{19.3}$$

for the inductance of a solenoid.

To obtain a result for the energy stored in an inductor, consider the charge dq passing through it in time dt: as dq crosses the inductor, the energy it loses is

$$dU = dq\,|\mathcal{E}| = dq\,L\,\frac{dI}{dt} = L\,\frac{dq}{dt}\,dI = LI\,dI$$

If we integrate this as the current through the inductor builds from zero to a final current I, we obtain for the total energy lost by the charge passing

[1] As noted in the previous chapter, an induced EMF is not a true voltage, but since in a circuit it behaves just like a true voltage, in this chapter we will simply refer to the EMFs across inductors as voltages.

Recall also that the dot over the I is not a Turkish letter; it is a shorthand notation for the time derivative:

$$\dot{I} = \frac{dI}{dt}$$

through the inductor

$$\begin{aligned} U &= \int dU \\ &= \int_0^I LI\, dI \\ &= \tfrac{1}{2}LI^2 \end{aligned} \tag{19.4}$$

This energy lost by the charge passing through the inductor is actually converted into a sort of magnetic potential energy: just like electric fields, magnetic fields have an energy density associated with them. In fact, if we use our result (19.3) for a solenoidal inductor and note that the volume of the solenoid is $A\ell$, the energy per unit volume associated with the magnetic field inside the solenoid is

$$\begin{aligned} u &= \frac{\frac{1}{2}LI^2}{A\ell} \\ &= \frac{\frac{1}{2}(\mu_0 n^2 A\ell)I^2}{A\ell} \\ &= \frac{1}{2\mu_0}(\mu_0 nI)^2 \\ &= \frac{1}{2\mu_0}B^2 \end{aligned} \tag{19.5}$$

Although we have derived this result only for a solenoid, it turns out to be completely general. The energy lost by the charge passing through the inductor is converted into the energy associated with the inductor's magnetic field: when the current through the inductor is increasing, passing through the inductor is like going uphill, with the kinetic energy of the charges being converted, not into gravitational, but into magnetic potential energy. And when the current through the inductor is decreasing, passing through the inductor is like going downhill, with magnetic potential energy being converted to kinetic energy.

Our unit for inductance is the Henry (H). Just so you know.

And while we're at it, the circuit symbol for an inductor, which is supposed to represent a coil of wire, is shown in fig. (19.1).

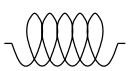


Figure 19.1: The Symbol for an Inductor

19.2 LR Circuits

LR circuits are very similar to RC circuits; you just have to be careful to get the signs right on the terms involving the inductance.[2]

First consider the case of an inductance L hooked up across a resistance R, with a current I_0 initially running through the inductor — what we will refer to as the case of current decay.[3] Applying the loop rule across the resistor and inductor, we have

$$-IR - L\frac{dI}{dt} = 0$$

where we have noted that if we go around the circuit with the current, we get a drop IR in voltage as we cross the resistor and a drop $L\dot{I}$ in voltage as we cross the inductor. Separating variables and integrating, we obtain

$$\begin{aligned} -\int_0^t \frac{R}{L}\,dt &= \int_{I_0}^I \frac{dI}{I} \\ -\frac{R}{L}\,t &= \ln\frac{I}{I_0} \\ I &= I_0\,e^{-tR/L} \end{aligned}$$

or

$$I = I_0\,e^{-t/\tau} \tag{19.6}$$

where $\tau = L/R$ is the time constant of the circuit, analogous to the time constant $\tau = RC$ for RC circuits.

Because the inductor limits the rate at which the current dies away, this exponential decay is just what we would have expected physically: At the

[2]It is always possible to get these signs right by careful reasoning, but you may find it cynically comforting to know if you do make a mistake, it should be obvious from your solution: inductors should always limit the rate at which change occurs, not cause runaway exponential currents.

[3]We might be in such a state as the result, for example, of having connected the battery in fig. (19.2) across the inductor and resistor by closing the switch and having reopened it when the current through the inductor had reached the value I_0.

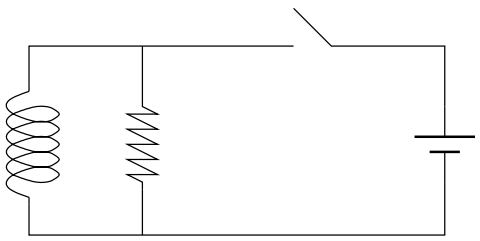


Figure 19.2: Death of a Current

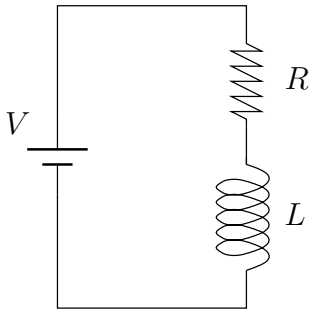


Figure 19.3: Birth of a Current

instant $t = 0$, the inductor maintains the current I_0, which, since this same current flows through the resistor, means a relatively large power loss $I_0^2 R$ in the resistor. And since energy being sapped by the resistor comes from the magnetic potential energy stored in the inductor, this also means a relatively rapid rate of decrease in the inductor's magnetic field and hence in the current through the inductor. But as the current dies away, so does the power loss in the resistor, with the result that the decay becomes less rapid and tails off.

Now consider the case, shown in fig. (19.3), of an inductance L hooked up across both a resistance R and a battery of voltage V, with no current initially running through the circuit — what we will refer to as the case of current buildup. Our loop equation will be the same as for the case of decay, just with an additional term for the battery's voltage:

$$V - IR - L\frac{dI}{dt} = 0$$

If we again separate variables and integrate, this yields

$$\begin{aligned} dt\,(V - IR) &= L\,dI \\ -R\,dt\left(I - \frac{V}{R}\right) &= L\,dI \\ -\int_0^t \frac{R}{L}\,dt &= \int_0^I \frac{dI}{I - \frac{V}{R}} \\ -\frac{R}{L}\,t &= \ln\left(\frac{I - \frac{V}{R}}{-\frac{V}{R}}\right) \\ I &= \frac{V}{R}\left(1 - e^{-tR/L}\right) \\ &= \frac{V}{R}\left(1 - e^{-t/\tau}\right) \end{aligned}$$

While the inductor again retards change by limiting the rate at which the current builds up, it is the resistor that determines the current's asymptotic

value V/R: as the current approaches this asymptotic value, the proportion of the battery's voltage V across the resistor increases and the leftover voltage difference $V - IR$ across the inductor decreases — with the result that the rate at which the current is building tails away and the current itself levels off.

19.3 Energy Density of the Electromagnetic Field

We will now derive our above results (16.24) and (19.5) for the energy densities of the electric and magnetic fields in a more rigorous way that will, as an added bonus,[4] also provide us with a result for the momentum densities of the fields.

Consider a charge dq, occupying a volume dV, that undergoes a displacement $d\mathbf{r}$ during time interval dt. As for any other force, the change in the electromagnetic potential energy and the work done by the electromagnetic force differ only in sign, so that $dU = -dW$: to the extent that the electric and magnetic fields do work on the charge, they lose energy. Since magnetic forces do no work, the change in the electromagnetic potential energy is attributable solely to the work done by the electric force $dq\,\mathbf{E}$:

$$dU = -dW = -d\mathbf{F} \cdot d\mathbf{r} = -dq\,\mathbf{E} \cdot d\mathbf{r}$$

The time rate of change of the electromagnetic field energy u per unit volume is thus

$$\frac{du}{dt} = \frac{dU}{dt\,dV} = \frac{-dq\,\mathbf{E} \cdot d\mathbf{r}}{dt\,dV} = -\frac{dq}{dV}\,\frac{d\mathbf{r}}{dt} \cdot \mathbf{E}$$

Since $d\mathbf{r}$ was the displacement of the charge dq, $d\mathbf{r}/dt$ is its velocity $\mathbf{v}$, and dq/dV is just the charge density ρ within the volume dV. Recalling that the current density $\mathbf{j}$ is by definition $\rho\mathbf{v}$, we therefore have

$$\frac{du}{dt} = -\rho\mathbf{v} \cdot \mathbf{E} = -\mathbf{j} \cdot \mathbf{E} \tag{19.7}$$

Because we are trying to arrive at a result for the electromagnetic field energy, we want to re-express this entirely in terms of electric and magnetic fields, and we can accomplish this by using Ampère's law

$$\nabla \times \mathbf{B} = \mu_0\,\mathbf{j} + \epsilon_0\mu_0\frac{\partial \mathbf{E}}{\partial t}$$

[4] Or so the infomercial marketing term goes, though how a bonus could be other than added is anyone's guess.

to solve for $\mathbf{j}$ in terms of $\mathbf{E}$ and $\mathbf{B}$:

$$\mathbf{j} = \frac{1}{\mu_0}\nabla \times \mathbf{B} - \epsilon_0 \frac{\partial \mathbf{E}}{\partial t}$$

Using this in eq. (19.7) yields

$$\begin{aligned} \frac{du}{dt} &= -\frac{1}{\mu_0}\mathbf{E}\cdot\nabla\times\mathbf{B} + \epsilon_0 \mathbf{E}\cdot\frac{\partial \mathbf{E}}{\partial t} \\ &= -\frac{1}{\mu_0}\mathbf{E}\cdot\nabla\times\mathbf{B} + \frac{\partial}{\partial t}\left(\tfrac{1}{2}\epsilon_0 E^2\right) \end{aligned} \tag{19.8}$$

where we have noted that, as for any vector,

$$\frac{\partial}{\partial t}E^2 = \frac{\partial}{\partial t}(\mathbf{E}\cdot\mathbf{E}) = 2\mathbf{E}\cdot\frac{\partial \mathbf{E}}{\partial t}$$

Our expression (19.8) is now entirely in terms of the electric and magnetic fields. We still, however, have some gerrymandering to do to get it into a form amenable to interpretation. With almost divine foresight, we will use Faraday's law

$$\nabla\times\mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t}$$

in the vector identity [5]

$$\nabla\cdot(\mathbf{E}\times\mathbf{B}) = \mathbf{B}\cdot\nabla\times\mathbf{E} - \mathbf{E}\cdot\nabla\times\mathbf{B}$$

to obtain

$$\begin{aligned} \mathbf{E}\cdot\nabla\times\mathbf{B} &= -\nabla\cdot(\mathbf{E}\times\mathbf{B}) + \mathbf{B}\cdot\nabla\times\mathbf{E} \\ &= -\nabla\cdot(\mathbf{E}\times\mathbf{B}) - \mathbf{B}\cdot\frac{\partial \mathbf{B}}{\partial t} \\ &= -\nabla\cdot(\mathbf{E}\times\mathbf{B}) - \frac{\partial}{\partial t}\left(\tfrac{1}{2}B^2\right) \end{aligned}$$

so that we can rewrite eq. (19.8) as

$$\begin{aligned} \frac{du}{dt} &= -\frac{1}{\mu_0}\left(-\mathbf{E}\cdot\nabla\times\mathbf{B} - \frac{\partial}{\partial t}\left(\tfrac{1}{2}B^2\right)\right) + \frac{\partial}{\partial t}\left(\tfrac{1}{2}\epsilon_0 E^2\right) \\ &= \nabla\cdot\left(\frac{1}{\mu_0}\mathbf{E}\times\mathbf{B}\right) + \frac{\partial}{\partial t}\left(\frac{1}{2\mu_0}B^2 + \frac{\epsilon_0}{2}E^2\right) \end{aligned} \tag{19.9}$$

[5] As you can verify be simply doing out the components, this identity of course holds for any two vectors, not just for $\mathbf{E}$ and $\mathbf{B}$.

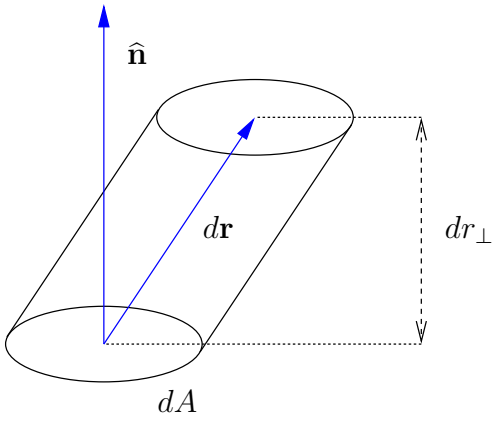


Figure 19.4: A Day in the Life of a Volume Element

Even if the route by which we have arrived at this isn't something you would have thought of on your own, you have to admit that eq. (19.9) is much more æsthetic than eq. (19.8).

Of course, that doesn't mean that interpreting eq. (19.9) is trivial, and in fact it will require some care. Recall that we are looking at a fixed charge dq, contained in a volume dV, that undergoes a displacement $d\mathbf{r}$ during time interval dt. That is, in everything we have done above our volume dV follows the charge dq, possibly also changing its shape and extent as it does so. What we want is a result for the energy density of the electromagnetic fields per unit *fixed* volume — within a volume element of a fixed shape and extent at a fixed location in our coordinate system. To relate our above result for a moving, morphing volume to a fixed one, let us distinguish the variable volume from fixed volumes by denoting the variable volume by ΔV and fixed volumes by dV. Now, since the charge dq contained within the volume ΔV does not change,

$$\frac{dq}{dt} = 0$$

which, if we use $dq = \rho\,\Delta V$, gives us

$$0 = \frac{dq}{dt} = \frac{d}{dt}(\rho\,\Delta V) = \frac{d\rho}{dt}\,\Delta V + \rho\,\frac{d(\Delta V)}{dt} \tag{19.10}$$

Now, the change $d(\Delta V)$ in the volume ΔV is due to the displacements $d\mathbf{r}$ of the points on its surface. In particular, the change in the volume associated with an infinitesimal patch dA of the surface of ΔV will, as you can see from fig. (19.4), be the area dA of the patch times the altitude of the cylinder formed by its infinitesimal displacement $d\mathbf{r}$, and this altitude in turn will be the normal component $dr_\perp = \hat{\mathbf{n}} \cdot d\mathbf{r}$ of the displacement $d\mathbf{r}$. The total change $d(\Delta V)$ in the volume ΔV will be the sum of these changes over all of the

patches dA that make up the surface S of ΔV:

$$d(\Delta V) = \oint_S dA\,\hat{\mathbf{n}} \cdot d\mathbf{r}$$

Since $d\mathbf{r}/dt$ is just the velocity $\mathbf{v}$ with which points on the surface of ΔV are moving, dividing this relation by dt to obtain the rate of change of ΔV yields

$$\frac{d(\Delta V)}{dt} = \oint_S dA\,\hat{\mathbf{n}} \cdot \frac{d\mathbf{r}}{dt} = \oint_S dA\,\hat{\mathbf{n}} \cdot \mathbf{v} \tag{19.11}$$

If we use Gauss's theorem, we can rewrite the right-hand side of this as

$$\oint_S dA\,\hat{\mathbf{n}} \cdot \mathbf{v} = \int dV\,\nabla \cdot \mathbf{v}$$

And since the volume ΔV is infinitesimal, there is only the contribution of the single volume element ΔV to this volume integral, so that eq. (19.11) reduces to

$$\frac{d(\Delta V)}{dt} = \int dV\,\nabla \cdot \mathbf{v} = \nabla \cdot \mathbf{v}\,\Delta V$$

Using this in eq. (19.10), we arrive at

$$0 = \frac{d\rho}{dt}\,\Delta V + \rho \nabla \cdot \mathbf{v}\,\Delta V$$

and hence, if we divide out the ΔV,

$$0 = \frac{d\rho}{dt} + \rho \nabla \cdot \mathbf{v} \tag{19.12}$$

Now, there are two kinds of contributions to the time rate of change of the charge density ρ: contributions from the explicit time-variation of ρ and contributions due to the change $d\mathbf{r}$ in the location of dq. The contributions from the explicit time-variation of ρ are given simply by

$$\frac{\partial \rho}{\partial t}$$

The contributions from the change in the location of dq are a bit more involved; these spatial contributions will be of the form

$$\begin{aligned} d\rho &= dx\,\frac{\partial \rho}{\partial x} + dy\,\frac{\partial \rho}{\partial y} + dz\,\frac{\partial \rho}{\partial z} \\ &= (dx\,\hat{\mathbf{x}} + dy\,\hat{\mathbf{y}} + dz\,\hat{\mathbf{z}}) \cdot \left(\frac{\partial \rho}{\partial x}\,\hat{\mathbf{x}} + \frac{\partial \rho}{\partial y}\,\hat{\mathbf{y}} + \frac{\partial \rho}{\partial z}\,\hat{\mathbf{z}} \right) \\ &= d\mathbf{r} \cdot \nabla \rho \end{aligned}$$

Thus to total rate of change of ρ, due both to explicit time-variation and to spatial displacements, is

$$\frac{d\rho}{dt} = \frac{\partial\rho}{\partial t} + \mathbf{v}\cdot\nabla\rho \tag{19.13}$$

Using eq. (19.13) in eq. (19.12), we have

$$0 = \frac{\partial\rho}{\partial t} + \mathbf{v}\cdot\nabla\rho + \rho\nabla\cdot\mathbf{v}$$

which, if we use the identity[6]

$$\nabla\cdot(a\mathbf{b}) = (\nabla a)\cdot\mathbf{b} + a\nabla\cdot\mathbf{b}$$

can be nicely rewritten as

$$0 = \frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\mathbf{v}) \tag{19.14}$$

Eq. (19.14) is known as the equation of continuity and applies to any kind of volume density within a fluid or region of continuous field, although the left-hand side vanishes only when the quantity in question (the charge dq in our above derivation) does not change: for the volume density ξ of any quantity, the total rate of change is of the form

$$\frac{\partial\xi}{\partial t} + \nabla\cdot(\xi\mathbf{v})$$

In particular, for the electromagnetic field energy density u we expect our rate of change (19.9) to be of the form

$$\frac{\partial u}{\partial t} + \nabla\cdot(u\mathbf{v}) = \nabla\cdot\left(\frac{1}{\mu_0}\mathbf{E}\times\mathbf{B}\right) + \frac{\partial}{\partial t}\left(\frac{1}{2\mu_0}B^2 + \frac{\epsilon_0}{2}E^2\right) \tag{19.15}$$

We therefore identify

$$u = \frac{1}{2\mu_0}B^2 + \frac{\epsilon_0}{2}E^2 \tag{19.16}$$

as the energy per unit volume of the electric and magnetic fields. And although we will not deal with it here, the quantity

$$\frac{1}{\mu_0}\mathbf{E}\times\mathbf{B}$$

in the other term is known as the Poynting vector and corresponds to the momentum density of the fields.

[6] You can easily verify this identity for yourself by expanding out both sides in terms of components and applying the product rule on the left-hand side.

19.4 Problems

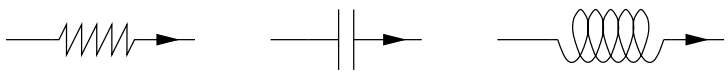


Figure 19.5: Problem 1

1. The currents through the resistor R, capacitor C, and inductor L in fig. (19.5) are to the right and increasing as time passes. For each circuit element, explain whether voltage is gained or dropped as you go through the element from left to right.

2. (a) Show that the equivalent inductance of inductors connected in series is given by
$$L_{eq} = \sum L_i$$
(b) Show that the equivalent inductance of inductors connected in parallel is given by
$$\frac{1}{L_{eq}} = \sum \frac{1}{L_i}$$

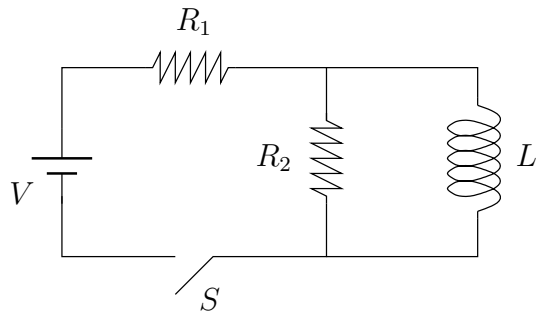


Figure 19.6: Problem 3

3. In the circuit shown in fig. (19.6) initially the switch S is open and nothing is happening. In temporal succession, you close switch S, wait a while, reopen switch S, wait a while, and then get bored and wander off. Determine

   i. The current in each branch of the circuit

   ii. The voltages across each resistor and the inductor

   at the following times:[7]

   (a) Immediately after switch S is closed.

   (b) A long time after switch S has been closed.

   (c) Immediately after switch S is reopened.

   (d) A long time after switch S has been reopened.

   (e) After you have lost interest and wandered off.

[7] Certain people — and we all know who *they* are — are bizarrely fascinated by circuits with switches, particularly circuits involving inductors. These people are obsessed with the phrases *immediately after* and *a long time later*, which is why we're using them here.

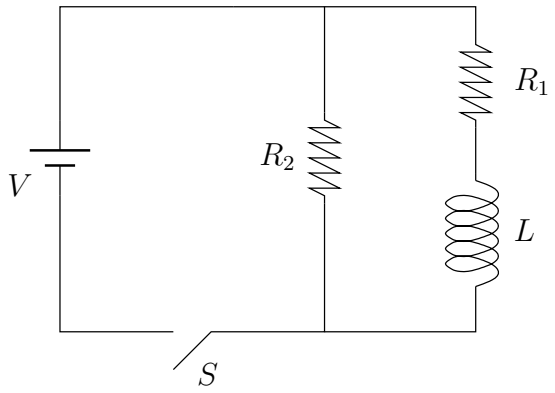


Figure 19.7: Problem 4

4. In the circuit shown in fig. (19.7) initially the switch S is open and nothing is happening. In temporal succession, you close switch S, wait a while, reopen switch S, wait a while, and then have a psychotic episode. Determine

   i. The current in each branch of the circuit

   ii. The time rate of change of the current in each branch of the circuit

   iii. The voltages across each resistor and the inductor

   at the following times:

   (a) Immediately after switch S is closed.

   (b) A long time after switch S has been closed.

   (c) Immediately after switch S is reopened.

   (d) Immediately after you smash the stupid circuit to pieces with a sledgehammer, pour gasoline on it, and light it on fire.

   (e) A long time after you have smashed the stupid circuit to pieces with a sledgehammer, poured gasoline on it, and lit it on fire.

5. (a) For the case of current decay in an LR circuit, show explicitly that

      i. At any given instant the power dissipated in the resistor equals the rate at which the inductor is losing energy.

      ii. The energy initially stored in the inductor equals the total energy dissipated in the resistor as the current completely dies away.

   (b) For the case of current buildup in an LR circuit, show explicitly that

      i. At any given instant power is conserved.

      ii. The total energy dissipated in the resistor and the energy ultimately stored in the inductor account for the total energy supplied by the battery as the current builds from zero to its asymptotic value.

6. A transformer is a device for stepping up an AC voltage to a higher value (or stepping it down to a lower value) by means of magnetic induction. It consists of two solenoids of the same length ℓ, one stuffed tightly inside the other, so that they have essentially the same cross sectional area A but differ in their total number of turns N_1 and N_2. The input solenoid is called the primary, the output solenoid the secondary.

   (a) Show that if the input is an AC voltage of amplitude V_{in}, then the output is an AC voltage of amplitude V_{out} given by

   $$\frac{V_{out}}{V_{in}} = \frac{N_{out}}{N_{in}}$$

   (b) How will the input power and current P_{in} and I_{in} be related to the output power and current P_{out} and I_{out}?

   (c) The power lines that convey electricity across the miles between the power plant and your home are high voltage — tens of thousands or even in excess of one hundred thousand volts — with transformers to step the voltage down to household 120 V at your end. The power company doesn't do this just to make it exciting when lines come down in storms: use $V = IR$ and $P = IV = I^2R = V^2/R$ to explain how high voltage in a power line that has a resistance of fixed value R leads to lower power loss in the line. See the footnote if you need a hint.[8]

7. Use our result $u = \frac{1}{2}\epsilon_0 E^2$ for the energy density of the electric field to explain why the charges that accumulate on the plates of a parallel-plate capacitor when it is connected to a battery are equal and opposite.

[8] Be careful not to confuse the voltage carried by the line with the voltage drop across the line's resistance.

19.5 Sketchy Answers

(1) Bwahahahaha! You can draw a definitive conclusion in only two of the three cases.

(3) The absolute values are

	#3a	#3b	#3c	#3d
I_{R_1}	$\frac{V}{R_1+R_2}$	$\frac{V}{R_1}$	0	0
I_{R_2}	$\frac{V}{R_1+R_2}$	0	$\frac{V}{R_1}$	0
I_L	0	$\frac{V}{R_1}$	$\frac{V}{R_1}$	0
V_{R_1}	$\frac{R_1}{R_1+R_2}V$	V	0	0
V_{R_2}	$\frac{R_2}{R_1+R_2}V$	0	$\frac{R_2}{R_1}V$	0
V_L	$\frac{R_2}{R_1+R_2}V$	0	$\frac{R_2}{R_1}V$	0

(4) The absolute values are

	#4a	#4b	#4c
I_{R_1}	0	$\dfrac{V}{R_1}$	$\dfrac{V}{R_1}$
I_{R_2}	$\dfrac{V}{R_2}$	$\dfrac{V}{R_2}$	$\dfrac{V}{R_1}$
I_L	0	$\dfrac{V}{R_1}$	$\dfrac{V}{R_1}$
$\dot{I}_{R_1}$	$\dfrac{V}{L}$	0	$\left(1+\dfrac{R_2}{R_1}\right)\dfrac{V}{L}$
$\dot{I}_{R_2}$	0	0	$\left(1+\dfrac{R_2}{R_1}\right)\dfrac{V}{L}$
$\dot{I}_L$	$\dfrac{V}{L}$	0	$\left(1+\dfrac{R_2}{R_1}\right)\dfrac{V}{L}$
V_{R_1}	0	V	V
V_{R_2}	V	V	$\dfrac{R_2}{R_1}V$
V_L	V	0	$\left(1+\dfrac{R_2}{R_1}\right)V$

Chapter 20

AC Circuits

20.1 Fourier Transforms

In AC circuits, the applied voltage is of the sinusoidal form [1]

$$V = V_0 \cos(\omega t + \varphi)$$

where V_0, ω, and φ are the amplitude, angular frequency, and phase of the voltage oscillation, respectively. Without loss of generality, we can immediately simplify this to

$$V = V_0 \cos \omega t$$

by setting up our time axis so that $\varphi = 0$. And since it's always nice to keep life simple, we will hereafter do so.

Although in an AC circuits the applied voltage, and hence the voltages across the various circuit elements, are now functions of time, at any given instant the same loop and junction rules that apply to DC circuits will apply to an AC circuit. For the AC *RC* circuit, for example, setting the sum of the applied voltage and the voltage drops across the resistor and capacitor to zero gives us [2]

$$V_{applied} + V_R + V_C = 0$$
$$V_0 \cos \omega t - IR - \frac{q}{C} = 0$$

[1] We could of course have equally well written this as

$$V = V_0 \sin(\omega t + \varphi)$$

Six of one, half a dozen of the other. Our choice of cosine rather than sine is dictated by no principle higher than cynical expedience.

[2] If you are a bit hazy about this, you might want to look back at the DC *RC* circuit in §16.6.

which, if we use

$$I = \frac{dq}{dt}$$

leaves us with a differential equation for q as a function of time:

$$V_0 \cos \omega t - \frac{dq}{dt} R - \frac{q}{C} = 0 \tag{20.1}$$

This equation may not look very nice, and the equations for more complicated AC circuits would be even worse: in general, circuits have multiple loops and junctions, which would leave us with coupled differential equations in multiple variables. Fortunately, there is a mathematical technique, the Fourier transform, that allows us to reduce such coupled differential equations to simultaneous algebraic equations. It turns out that any reasonably well-behaved function can be written as a superposition of sine and cosine waves — or, equivalently and more æsthetically, as a superposition of complex exponentials of the form [3]

$$e^{i\omega t} = \cos \omega t + i \sin \omega t$$

each of which is, so to speak, like a sine wave and a cosine wave rolled together. To reproduce a function $g(t)$, we need to superpose the contributions of such exponentials for all possible values of ω, from $-\infty$ to $+\infty$, in just the right proportions:

$$g(t) = \int_{-\infty}^{\infty} d\omega \, e^{i\omega t} \, \tilde{g}(\omega) \tag{20.2}$$

where the factor $\tilde{g}(\omega)$ ensures that the contribution of $e^{i\omega t}$ has the required weight in the overall superposition. In other words, $\tilde{g}(\omega)$ is the amplitude of the wave of angular frequency ω in the superposition for $g(t)$. This animal $\tilde{g}(\omega)$ is called the Fourier transform of $g(t)$.

The immediate issue is determining the required $\tilde{g}(\omega)$ in eq. (20.2). If, with the usual divine foresight, we multiply both sides of eq. (20.2) by $e^{-i\omega' t}$ and integrate over t, we have

$$\begin{aligned} \int_{-\infty}^{\infty} dt \, e^{-i\omega' t} \, g(t) &= \int_{-\infty}^{\infty} dt \, e^{-i\omega' t} \int_{-\infty}^{\infty} d\omega \, e^{i\omega t} \, \tilde{g}(\omega) \\ &= \int_{-\infty}^{\infty} d\omega \, \tilde{g}(\omega) \int_{-\infty}^{\infty} dt \, e^{i(\omega - \omega')t} \end{aligned} \tag{20.3}$$

[3] Although we will be working through the details of Fourier transforms and the calculations for AC circuits explicitly, we assume that you already have a familiarity and facility with the basics of complex numbers. Those of you who are a bit rusty with them will find a brief refresher on p.974.

This might not seem like progress, but as we will demonstrate in §20.3, the inner integral yields

$$\int_{-\infty}^{\infty} dt\, e^{i(\omega-\omega')t} = 2\pi\delta(\omega-\omega')$$

where $\delta(\omega-\omega')$ — known, by a strange coincidence, as the delta function — has the odd properties that

- $\delta(\omega-\omega') = 0$ when $\omega \neq \omega'$.
- When $\omega = \omega'$, $\delta(\omega-\omega')$ is an infinite spike such that

$$\int d\omega\, \delta(\omega-\omega') = 1$$

  over any range containing $\omega = \omega'$.

Thus the delta function $\delta(\omega-\omega')$ is just a "blip" at $\omega = \omega'$ and has the effect that any integrand containing it ends up being evaluated at $\omega = \omega'$. That is, for any function $\varphi(\omega)$,

$$\int_{-\infty}^{\infty} d\omega\, \varphi(\omega)\, \delta(\omega-\omega') = \varphi(\omega')$$

Using these properties of the delta function reduces eq. (20.3) to

$$\int_{-\infty}^{\infty} dt\, e^{-i\omega' t}\, g(t) = \int_{-\infty}^{\infty} d\omega\, \tilde{g}(\omega)\, 2\pi\delta(\omega-\omega') = 2\pi\tilde{g}(\omega')$$

so that we obtain as the general expression for the Fourier component $\tilde{g}(\omega)$ of the function $g(t)$

$$\tilde{g}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dt\, g(t)\, e^{-i\omega t} \tag{20.4}$$

It is worth noting the nice "symmetry" in eqq. (20.2) and (20.4):

$$g(t) = \int_{-\infty}^{\infty} d\omega\, e^{i\omega t}\, \tilde{g}(\omega) \qquad\qquad \tilde{g}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dt\, g(t)\, e^{-i\omega t}$$

To get $g(t)$, you integrate $\tilde{g}(\omega)$ with an $e^{i\omega t}$; to get $\tilde{g}(\omega)$ you integrate $g(t)$ with an $e^{-i\omega t}$. Of course, the $1/2\pi$ occurs in a lopsided way, but even that could be made symmetric: as you will no doubt find in your further reading, different people divvy up this $1/2\pi$ differently; since all that matters is that together the transform and inverse-transform relations (20.2) and (20.4) have an overall factor of $1/2\pi$, you can put the whole $1/2\pi$ with the integral of $g(t)$ (as we have done), or you can put it all with the $\tilde{g}(\omega)$, or you can split it symmetrically so that the integral in each relation has a factor of $1/\sqrt{2\pi}$. [4]

[4] If you enjoy being ironically whimsical or are socially maladjusted, you could of course also opt to put $1/(2\pi)^{\frac{2}{3}}$ with one integral and $1/(2\pi)^{\frac{1}{3}}$ with the other, or $1/(2\pi)^3$ with one and $(2\pi)^2$ with the other, or any of an infinite number of equivalent schemes. Sometimes mathematics provides more scope for making personal statements than you would expect.

Anyway, Fourier transforms will reduce our differential equations for AC circuits to algebraic equations. For example, for the equation governing the AC RC circuit, eq. (20.1), we first rewrite the applied voltage $V = V_0 \cos \omega t$ in the form

$$V = V_0 \, e^{i\omega t} \tag{20.5}$$

with it being understood that the physical applied voltage is the real part of this complex expression. If we then write the solution for $q(t)$ in terms of its Fourier transform $\widetilde{q}(\omega')$ by [5]

$$q(t) = \int_{-\infty}^{\infty} d\omega' \, e^{i\omega' t} \, \widetilde{q}(\omega') \tag{20.6}$$

we have

$$\begin{aligned} I &= \frac{dq}{dt} \\ &= \frac{d}{dt} \int_{-\infty}^{\infty} d\omega' \, e^{i\omega' t} \, \widetilde{q}(\omega') \\ &= \int_{-\infty}^{\infty} d\omega' \, i\omega' e^{i\omega' t} \, \widetilde{q}(\omega') \end{aligned} \tag{20.7}$$

Here we can start to see already how Fourier transforms change differential equations into algebraic equations: under the Fourier transform, what was a time derivative has by virtue of

$$\frac{d}{dt} \, e^{i\omega' t} = i\omega' e^{i\omega' t}$$

become merely an algebraic factor of $i\omega'$. And in an exactly similar way, integrals with respect to time would yield factors of $1/i\omega'$:

$$\int dt \, e^{i\omega' t} = \frac{1}{i\omega'} \, e^{i\omega' t}$$

But to return to the business at hand, using eqq. (20.5), (20.6), and (20.7) in eq. (20.1) gives

$$V_0 \, e^{i\omega t} - R \int_{-\infty}^{\infty} d\omega' \, i\omega' e^{i\omega' t} \, \widetilde{q}(\omega') - \frac{1}{C} \int_{-\infty}^{\infty} d\omega' \, e^{i\omega' t} \, \widetilde{q}(\omega') = 0 \tag{20.8}$$

In order to combine the $V_0 \, e^{i\omega t}$ term with the other two on the left-hand side, we would like to rewrite $V_0 \, e^{i\omega t}$ in terms of an integral by $d\omega'$, and we can do this by means of a delta function:

$$V_0 \, e^{i\omega t} = \int_{-\infty}^{\infty} d\omega' \, V_0 \, e^{i\omega' t} \delta(\omega' - \omega)$$

[5] Since ω has already been used for the angular frequency of the applied voltage, to avoid confusion we use ω' for the dummy variable of integration.

Using this in eq. (20.8) and combining terms, we arrive at

$$\int_{-\infty}^{\infty} d\omega'\, e^{i\omega' t}\Big(V_0\,\delta(\omega'-\omega) - i\omega' R\,\widetilde{q}(\omega') - \frac{1}{C}\,\widetilde{q}(\omega')\Big) = 0 \tag{20.9}$$

Since the various $e^{i\omega' t}$ over which we are integrating are functionally independent of each other, in order for this integral to vanish for all times t the quantity in parentheses must vanish: [6]

$$V_0\,\delta(\omega'-\omega) - i\omega'\widetilde{q}(\omega')R - \frac{1}{C}\,\widetilde{q}(\omega') = 0$$

The Fourier transform has, as advertised, reduced the differential eq. (20.1) to an algebraic equation. If we multiply both sides by C and combine the $\widetilde{q}$ terms, we have

$$CV_0\,\delta(\omega'-\omega) - (1 + i\omega' RC)\,\widetilde{q}(\omega') = 0 \tag{20.10}$$

As long as $1 + i\omega' RC \neq 0$ — a condition into which we will look in more detail toward the end of this section — the solution for $\widetilde{q}$ is

$$\widetilde{q}(\omega') = \frac{CV_0}{1 + i\omega' RC}\,\delta(\omega'-\omega) \tag{20.11}$$

Plugging this back into eq. (20.6) then gives us

$$\begin{aligned} q(t) &= \int_{-\infty}^{\infty} d\omega'\, e^{i\omega' t}\,\widetilde{q}(\omega') \\ &= \int_{-\infty}^{\infty} d\omega'\, e^{i\omega' t}\,\frac{CV_0}{1 + i\omega' RC}\,\delta(\omega'-\omega) \\ &= e^{i\omega t}\,\frac{CV_0}{1 + i\omega RC} \end{aligned}$$

[6] If you aren't convinced by this argument, you can just take the inverse Fourier transform of eq. (20.9):

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} d\omega''\, e^{-i\omega'' t} \int_{-\infty}^{\infty} d\omega'\, e^{i\omega' t}\big(V_0\,\delta(\omega'-\omega) - i\omega' R\,\widetilde{q}(\omega') - \frac{1}{C}\,\widetilde{q}(\omega')\big) \\ &= \int_{-\infty}^{\infty} d\omega'\big(V_0\,\delta(\omega'-\omega) - i\omega' R\,\widetilde{q}(\omega') - \frac{1}{C}\,\widetilde{q}(\omega')\big) \int_{-\infty}^{\infty} d\omega''\, e^{i(\omega'-\omega'')t} \\ &= \int_{-\infty}^{\infty} d\omega'\big(V_0\,\delta(\omega'-\omega) - i\omega' R\,\widetilde{q}(\omega') - \frac{1}{C}\,\widetilde{q}(\omega')\big)\, 2\pi\,\delta(\omega'-\omega'') \\ &= 2\pi\big(V_0\,\delta(\omega''-\omega) - i\omega'' R\,\widetilde{q}(\omega'') - \frac{1}{C}\,\widetilde{q}(\omega'')\big) \end{aligned}$$

Alternatively (but equivalently), a comparison of eqq. (20.2) and (20.9) shows that the quantity in parentheses represents the Fourier transform $\widetilde{g}(\omega')$ of a function $g(t) = 0$. And by eq. (20.4), if $g(t)$ vanishes, so does $\widetilde{g}(\omega')$.

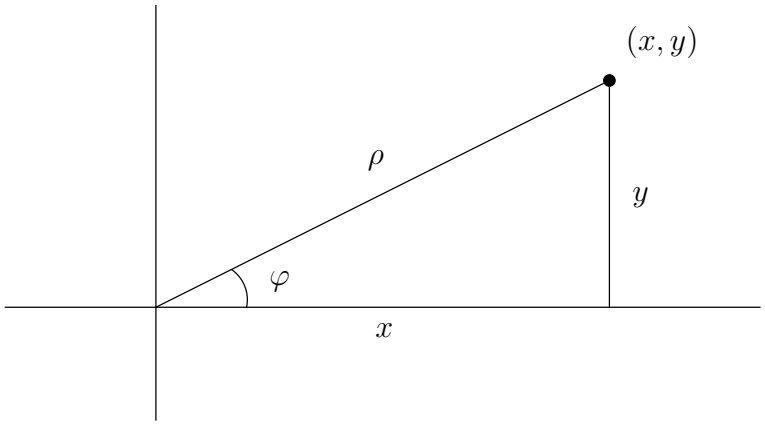


Figure 20.1: The Complex Plane

$$= \frac{CV_0\, e^{i\omega t}}{1 + i\omega RC}$$
$$= \frac{CV}{1 + i\omega RC}$$

where in the last step we have used eq. (20.5).

From this solution for $q(t)$, we can obtain results for the various other quantities in the circuit. The voltage V_C across the capacitor as a function of time, for example, will be

$$V_C = \frac{q}{C} = \frac{V}{1 + i\omega RC} \tag{20.12}$$

The only catch is that we have to remember that the physical voltage

$$V = V_0 \cos \omega t$$

was only the real part of the expression $V_0 e^{i\omega t}$ we used for the voltage in eq. (20.8), so that the physical values of $q(t)$ and any quantities we derive from $q(t)$ will likewise be only their real parts. Since in circuits we are usually concerned with the amplitudes and phases of these quantities, it is most convenient to rewrite them in polar form when we are extracting their real parts — which means that this would be a good time to offer a little refresher for those of you who are a bit rusty at manipulating complex numbers. Those of you whose facility with complex numbers hasn't oxidized may skip over this and resume below at the heading "Back to Business."

A Brief Refresher in Complex Arithmetic

Recall that any complex number $z = x + iy$ may be regarded as a point (x, y) in the complex plane. As you can see from fig. (20.1), the relations between

the Cartesian and polar coordinates ρ and φ of a point (x, y) are

$$x = \rho\cos\varphi \qquad\qquad y = \rho\sin\varphi$$
$$\rho = \sqrt{x^2 + y^2} \qquad\qquad \tan\varphi = \frac{y}{x}$$

By means of the Euler relation

$$e^{i\varphi} = \cos\varphi + i\sin\varphi$$

the complex number $z = x + iy$ may therefore also be expressed in the equivalent polar form $z = \rho e^{i\varphi}$:

$$z = \rho e^{i\varphi} = \rho(\cos\varphi + i\sin\varphi) = \rho\cos\varphi + i\rho\sin\varphi = x + iy \quad \checkmark$$

The angle φ is called the phase.

The complex conjugate of z, usually denoted by z^* or $\bar{z}$, is

$$z^* = x - iy$$

or, equivalently,

$$z^* = \rho e^{-i\varphi}$$

as you can see by

$$\rho e^{-i\varphi} = \rho\Big(\cos\varphi - i\sin\varphi) = \rho\cos\varphi - i\rho\sin\varphi = x - iy \quad \checkmark$$

Note in particular that

$$zz^* = (x + iy)(x - iy) = x^2 - i^2y^2 = x^2 + y^2 = \rho^2$$

It is usually easier to see what's going on when the imaginary contributions to a complex number are entirely in the numerator. To accomplish this when confronted with something like

$$\frac{1}{x + iy}$$

the trick is

$$\frac{1}{x + iy} = \frac{1}{x + iy}\,\frac{x - iy}{x - iy} = \frac{x - iy}{x^2 + y^2}$$

And, trivial as it may seem, it is also useful to note that if

$$z = \frac{a + ib}{C}$$

where C is real, then

$$\tan\varphi = \frac{b/C}{a/C} = \frac{b}{a}$$

Also frequently useful are

$$\frac{1}{i} = -i$$

and

$$\begin{aligned} e^{i\frac{\pi}{2}} &= \cos\frac{\pi}{2} + i\sin\frac{\pi}{2} = i \\ e^{-i\frac{\pi}{2}} &= \cos\frac{\pi}{2} - i\sin\frac{\pi}{2} = -i \\ e^{i\pi} &= \cos\pi + i\sin\pi \ = -1 \end{aligned}$$

Back to Business

Now back to eq. (20.12),

$$V_C = \frac{q}{C} = \frac{V}{1+i\omega RC} = \frac{V_0}{1+i\omega RC}\, e^{i\omega t} \tag{20.12}$$

To extract the real and thus physical part of the voltage V_C across the capacitor, we first move the imaginary parts into the numerator by noting that

$$\frac{1}{1+i\omega RC} = \frac{1}{1+i\omega RC}\,\frac{1-i\omega RC}{1-i\omega RC} = \frac{1-i\omega RC}{1+(\omega RC)^2}$$

This fraction may therefore be written in polar form with

$$\begin{aligned} \rho &= \sqrt{\left(\frac{1}{1+i\omega RC}\right)\left(\frac{1}{1+i\omega RC}\right)^*} \\ &= \frac{1}{\sqrt{1+(\omega RC)^2}} \\ \tan\varphi &= \frac{-\omega RC}{1} \\ &= -\omega RC \end{aligned}$$

Thus

$$V_C = \frac{V_0}{\sqrt{1+(\omega RC)^2}}\,\exp\Big(i\big(\omega t - \tan^{-1}(\omega RC)\big)\Big)$$

The physical voltage across the capacitor is the real part of this:

$$V_{C,\,physical} = \frac{V_0}{\sqrt{1+(\omega RC)^2}}\,\cos\Big(\omega t - \tan^{-1}(\omega RC)\Big) \tag{20.13}$$

From eq. (20.13) we see that the amplitude or peak value of the voltage across the capacitor differs from that of the applied voltage by a frequency-dependent factor of

$$\frac{1}{\sqrt{1+(\omega RC)^2}} \tag{20.14}$$

The peak voltage across the capacitor is therefore small when the voltage driving the RC circuit has an angular frequency ω large compared to $1/RC$. If we recall that RC was the time constant of the DC RC circuit — it gave a measure of how long it took the DC RC circuit to charge or discharge — , this makes physical sense: the factor (20.14) is telling us that when the frequency of the driving voltage is high compared to the time scale over which the RC circuit can charge and discharge, during each cycle there isn't enough time to accumulate much charge and hence voltage across the capacitor. At the other extreme, as $\omega \to 0$, the factor (20.14) becomes unity: the applied voltage ceases oscillating and becomes a DC voltage, with the result that the capacitor has all the time in the world to charge up to the full applied voltage.

We also see from eq. (20.13) that the voltage across the capacitor has its peak an angle of $\tan^{-1}(\omega RC)$ *after* the applied voltage has its peak: since the applied voltage goes as $\cos\omega t$ and the voltage across the capacitor as

$$\cos\Big(\omega t - \tan^{-1}(\omega RC)\Big)$$

the angle $\omega t - \tan^{-1}(\omega RC)$ for the capacitor does not reach the same value as that for the applied voltage until a somewhat later t. This is another consequence of the delay in charging and discharging the capacitor: it takes time for charge to flow onto the capacitor in response to the applied voltage, with the result that the charge and voltage across the capacitor do not reach their peak values until some time after the peak in the applied voltage. And the magnitude of this effect once again depends on the angular frequency of the driving voltage versus the time constant RC of the RC circuit, in the same form ωRC.

The most important point, however, is that $q(t)$, and hence all time-dependent quantities in the circuit, oscillate at the same angular frequency ω as the driving voltage V, regardless of the values of R and C. And this effect is far more general than the RC circuit: ultimately it came about because of the $\delta(\omega' - \omega)$ in the term for the driving voltage, which will be common to all AC circuits. As we will see in the next section, knowing that an AC circuit will respond at the same angular frequency as the driving voltage will enable us to analyze it by an easier alternative method. We will therefore abandon our present analysis in favor of developing this alternative method.

First, however, we need to tie up one loose end: In eq. (20.10),

$$CV_0\,\delta(\omega'-\omega)-(1+i\omega' RC)\,\widetilde{q}(\omega')=0$$

we had assumed that $1+i\omega' RC\neq 0$ when we solved for $\widetilde{q}(\omega')$. To see what happens in the special case $1+i\omega' RC=0$, note that the terms

$$(1+i\omega' RC)\,\widetilde{q}(\omega')$$

in this equation originally came from the terms

$$-\frac{dq}{dt}R-\frac{q}{C}$$

in eq. (20.1). The case $1+i\omega' RC=0$ is therefore equivalent to

$$-\frac{dq}{dt}R-\frac{q}{C}=0 \tag{20.15}$$

As you may recall from §16.6, eq. (20.15) is precisely the relation for a discharging DC RC circuit and has a solution of the form

$$q=q_0e^{-t/RC} \tag{20.16}$$

The case $1+i\omega' RC=0$ thus corresponds physically to an exponentially decaying initial charge on the capacitor. Solutions that die away like this are called transients, and the general solution to our AC RC circuit, and indeed to any AC circuit, is actually a superposition of a perpetual oscillatory solution like (20.13) and a transient like (20.16). Having pointed this out, we will, however, hereafter neglect such transients: in practice they die away quickly, and at any rate our concern is rather with the longterm, steady-state behavior of AC circuits.

20.2 Impedance

The method of Fourier transforms is much easier than directly solving the coupled differential equations of AC circuits, but there is a still easier, softer way. Since the response of a circuit will always be at the same angular frequency ω as the driving voltage applied to it, the current flowing through or into any circuit element must be of the form

$$I=I_0\,e^{i(\omega t+\varphi)} \tag{20.17}$$

That is, the current I may differ from the driving voltage in its phase φ, but it will always share the same angular frequency ω of oscillation.

For a resistor, the voltage V across the resistor is related to the current I through it by Ohm's bogus law, $V = IR$. For a voltage of the form

$$V = V_0\, e^{i\omega t} \tag{20.18}$$

the current is therefore

$$I = \frac{V}{R} = \frac{V_0}{R}\, e^{i\omega t}$$

Comparing this with the general result (20.17), we see that we have

$$I_0\, e^{i(\omega t+\varphi)} = \frac{V_0}{R}\, e^{i\omega t}$$

and hence

$$I_0 = \frac{V_0}{R} \qquad\qquad \varphi = 0 \tag{20.19}$$

That is, the current through the resistor is in phase with the voltage across it — so that they both have their peaks at the same time — , and the amplitudes of the voltage and current are related simply by Ohm's bogus law: there is no difference between the DC and AC behavior of the resistor.

For a capacitor, $q = CV$. To relate the voltage to the *current* and thus obtain something that looks like Ohm's bogus law, we need merely differentiate $q = CV$ with respect to time and use eqq. (20.17) and (20.18):

$$\begin{aligned}
\frac{dq}{dt} &= C\,\frac{dV}{dt} \\
I &= C\,\frac{d}{dt}\Big(V_0\, e^{i\omega t}\Big) \\
I_0\, e^{i\omega t+\varphi} &= i\omega C V_0\, e^{i\omega t} \\
&= e^{i\frac{\pi}{2}}\,\omega C V_0\, e^{i\omega t} \\
&= \omega C V_0\, e^{i(\omega t+\frac{\pi}{2})}
\end{aligned}$$

From this we see that

$$I_0 = \omega C V_0 \qquad\qquad \varphi = +\frac{\pi}{2} \tag{20.20}$$

Thus the current flowing into a capacitor *leads* the voltage across it by a quarter cycle. Physically, we would expect the current to lead the voltage: you get a voltage across a capacitor only to the extent that you have charge across it, and there won't be a charge across it until current has flowed into it.

For an inductor, we have to be careful about our signs. If we are going to express the voltage across the inductor in a way that is consistent with

Ohm's bogus law, which is expressed in terms of the voltage *drop* across the resistor, we want to use $+L\dot{I}$ rather than $-L\dot{I}$. Then eqq. (20.17) and (20.18) give us

$$\begin{aligned} V &= L\frac{dI}{dt} \\ V_0\, e^{i\omega t} &= L\frac{d}{dt} I_0\, e^{i(\omega t+\varphi)} \\ &= i\omega L I_0\, e^{i(\omega t+\varphi)} \\ &= e^{i\frac{\pi}{2}}\, \omega L I_0\, e^{i(\omega t+\varphi)} \\ &= I_0 \omega L\, e^{i(\omega t+\varphi+\frac{\pi}{2})} \end{aligned}$$

From this, we see that

$$I_0 = \frac{V_0}{\omega L} \qquad\qquad \varphi = -\frac{\pi}{2} \tag{20.21}$$

That is, the current through an inductor *lags* the voltage applied across it by a quarter cycle. Physically, we would expect the current to lag the voltage: the inductor doesn't like change, so that there will be a time lag between the application of a voltage across the inductor and the flow of current through it.

If we write the above current-voltage relations for resistors, capacitors, and inductors in the forms

$$V = IR \qquad\qquad V = I\,\frac{1}{i\omega C} \qquad\qquad V = I\, i\omega L$$

we see that we have arrived at an AC version

$$V = IZ$$

of Ohm's bogus law, where the impedance Z, the AC equivalent of resistance, is

$$Z = \begin{cases} R & \text{for a resistor} \\ \dfrac{1}{i\omega C} & \text{for a capacitor} \\ i\omega L & \text{for an inductor} \end{cases} \tag{20.22}$$

So instead of using Fourier transforms to solve a differential equation for each circuit, we can simply extend the DC relation $V = IR$ and the rules for combining resistors in series and parallel to AC circuits by using the impedances of eq. (20.22) in $V = IZ$ — a technique we will now apply to the RC, RL, and RLC circuits.

20.2.1 The RC Circuit

We have of course already worked out the AC RC circuit by the method of Fourier transform at the beginning of this chapter, but we will here analyze it anew as an exercise in working with impedances and the AC version of Ohm's bogus law.

For a simple RC series circuit driven by an AC voltage, we have an equivalent impedance of

$$Z_{eq} = Z_R + Z_C = R + \frac{1}{i\omega C} = R\Big(1 + \frac{1}{i\omega RC}\Big) = R\Big(1 - i\,\frac{1}{\omega RC}\Big)$$

As in our earlier analysis, it will be convenient to express this in the polar form $\rho e^{i\varphi}$:

$$\rho = R\sqrt{1 + \frac{1}{(\omega RC)^2}} \qquad\qquad \tan\varphi = -\frac{1}{\omega RC}$$

so that

$$Z_{eq} = R\sqrt{1 + \frac{1}{(\omega RC)^2}}\ \exp\Big(-i\tan^{-1}\frac{1}{\omega RC}\Big)$$

For the current flowing through the circuit, the AC version of Ohm's bogus law therefore gives

$$\begin{aligned} I &= \frac{V}{Z_{eq}} \\ &= \frac{V}{R\sqrt{1 + \frac{1}{(\omega RC)^2}}}\ \exp\Big(i\tan^{-1}\frac{1}{\omega RC}\Big) \\ &= \frac{V}{R}\frac{1}{\sqrt{1 + \frac{1}{(\omega RC)^2}}}\ \exp\Big(i\tan^{-1}\frac{1}{\omega RC}\Big) \end{aligned} \tag{20.23}$$

The current thus leads the applied voltage by an angle of

$$\tan^{-1}\frac{1}{\omega RC}$$

The voltage across the resistor, which is similarly given by the AC version of Ohm's bogus law as $V_R = IZ_R = IR$, is then

$$V_R = \frac{V}{\sqrt{1 + \frac{1}{(\omega RC)^2}}}\ \exp\Big(i\tan^{-1}\frac{1}{\omega RC}\Big)$$

Thus the ratio of the peak value of the voltage across the resistor to the peak value of the voltage applied to the circuit is

$$\frac{V_{0R}}{V_0} = \frac{1}{\sqrt{1 + \frac{1}{(\omega RC)^2}}} \tag{20.24}$$

and the voltage across the resistor leads the applied voltage by the same angle of

$$\tan^{-1}\frac{1}{\omega RC}$$

as the current.

Finally, the voltage across the capacitor is given by the AC version of Ohm's bogus law as

$$\begin{aligned}
V_C &= IZ_C \\
&= I\frac{1}{i\omega C} \\
&= \left(\frac{V}{R}\frac{1}{\sqrt{1+\frac{1}{(\omega RC)^2}}}\ \exp\left(i\tan^{-1}\frac{1}{\omega RC}\right)\right)\frac{1}{i\omega C} \\
&= \frac{1}{i}\frac{V}{\omega RC\sqrt{1+\frac{1}{(\omega RC)^2}}}\ \exp\left(i\tan^{-1}\frac{1}{\omega RC}\right) \\
&= e^{-i\frac{\pi}{2}}\frac{V}{\sqrt{1+(\omega RC)^2}}\ \exp\left(i\tan^{-1}\frac{1}{\omega RC}\right) \\
&= \frac{V}{\sqrt{1+(\omega RC)^2}}\ \exp\left(-i\left(\frac{\pi}{2}-\tan^{-1}\frac{1}{\omega RC}\right)\right) \\
&= \frac{V}{\sqrt{1+(\omega RC)^2}}\ e^{-i\tan^{-1}(\omega RC)}
\end{aligned}$$

where in the last step we have noted that if

$$\varphi = \tan^{-1}\frac{1}{\omega RC}$$

then

$$\tan\left(\frac{\pi}{2}-\varphi\right) = \cot\varphi = \frac{1}{\tan\varphi} = \frac{1}{1/\omega RC} = \omega RC$$

so that

$$\frac{\pi}{2}-\varphi = \tan^{-1}(\omega RC)$$

or in other words

$$\frac{\pi}{2}-\tan^{-1}\frac{1}{\omega RC} = \tan^{-1}(\omega RC)$$

The ratio of peak value of the voltage across the capacitor to the peak value of the applied voltage is thus

$$\frac{V_{0C}}{V_0} = \frac{1}{\sqrt{1+(\omega RC)^2}} \tag{20.25}$$

and the voltage across the capacitor lags the applied voltage by an angle of

$$\tan^{-1}(\omega RC)$$

And now it's time to interpret this whole mess.

First note that everything in the circuit depends on the value of the driving angular frequency ω and on the values of the resistance R and capacitance C only through the combination ωRC: the behavior of the circuit depends entirely on how rapidly the circuit is driven in comparison to its time constant RC. Figg. (20.2) and (20.3) show the plots of eqq. (20.24) and (20.25): the red line is V_{0R}/V_0 and the blue V_{0C}/V_0. As you can see both from eqq. (20.24) and (20.25) and from the plots, at high frequencies ($\omega RC \gg 1$),

$$\frac{V_{0R}}{V_0} \to 1 \qquad\qquad \frac{V_{0C}}{V_0} \to 0$$

This is what we would expect: at high frequencies, there isn't enough time for charge to accumulate across the capacitor, so the entire applied voltage is across the resistor. At low frequencies ($\omega RC \ll 1$), the situation is just the opposite:

$$\frac{V_{0R}}{V_0} \to 0 \qquad\qquad \frac{V_{0C}}{V_0} \to 1$$

This is again what we would expect: at low frequencies, the capacitor has ample time to fully charge and discharge, so that the applied voltage ends up being entirely across the capacitor. Note also that, as is most visible at intermediate frequencies $\omega RC \sim 1$,

$$\frac{V_{0R}}{V_0} + \frac{V_{0C}}{V_0} \neq 1$$

because the peaks in the voltages across the resistor and capacitor are out of phase with each other.

The AC RC circuit is sometimes called an *RC* filter: If the applied voltage is the input and the output is the voltage across the resistor, then there is very little output for — that is, very little transmission of — the low-frequency components of the input signal but very nearly lossless output for the high-frequency components. When the voltage across the resistor is used as the output, the AC RC circuit is therefore called a high-pass filter. Since whatever portion of the applied voltage that is not across the resistor is across the capacitor, the effect is just the opposite when the voltage across the capacitor is used as the output: there is very little transmission of the high-frequency components of the input signal but very nearly lossless output for the low-frequency components, and in this case the AC RC circuit is called a low-pass filter. By adjusting the value of RC, one can adjust the ranges that constitute high and low frequencies.

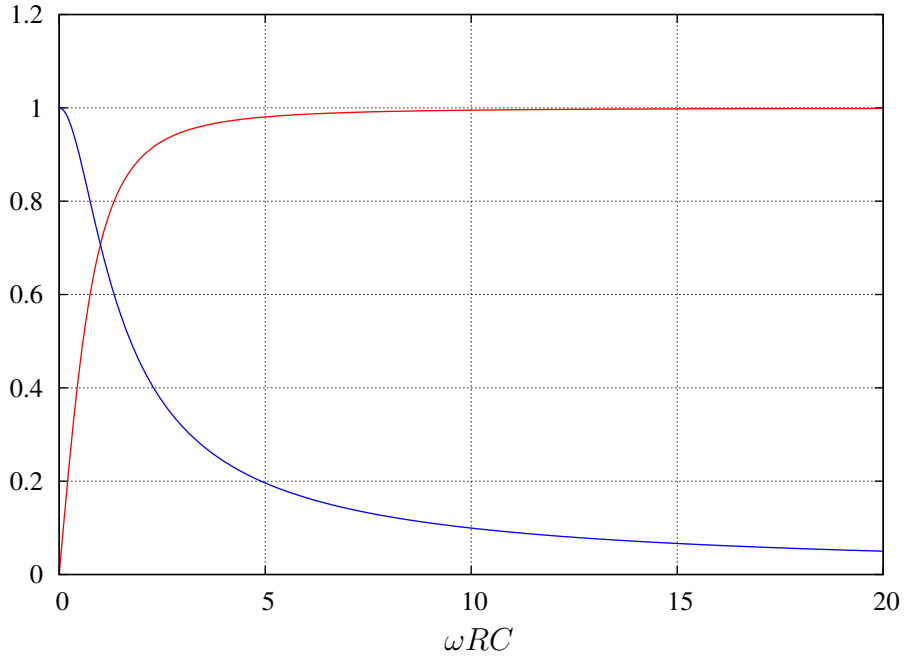


Figure 20.2: High-Frequency Behavior of an RC Circuit

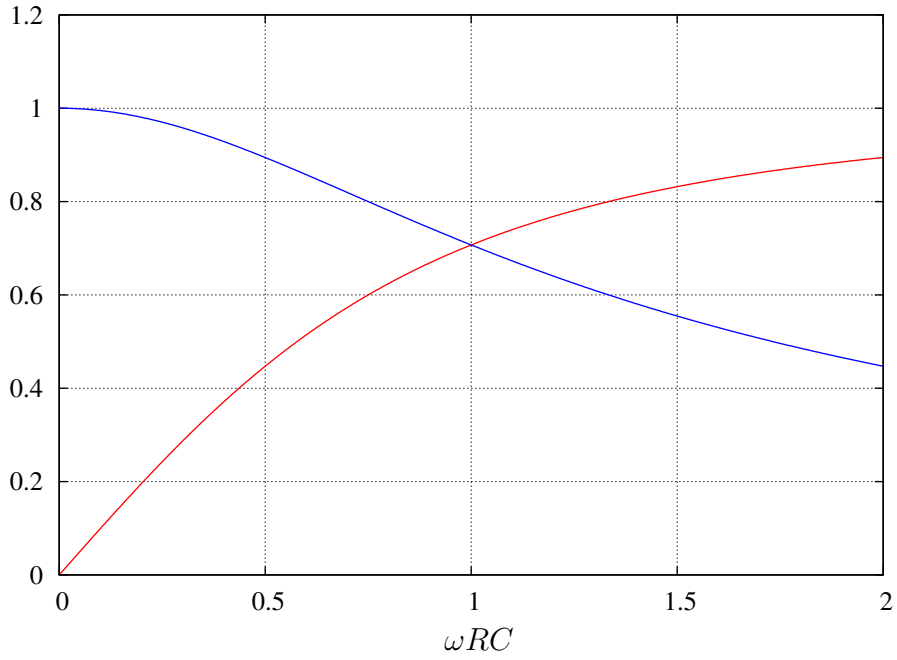


Figure 20.3: Low-Frequency Behavior of an RC Circuit

20.2.2 The LR Circuit

The analysis of the series LR circuit is virtually identical to that of the series RC circuit, so we won't go through the details of the calculation here. You should find that the equivalent impedance of the circuit is

$$Z_{eq} = R + i\omega L = \sqrt{R^2 + (\omega L)^2}\ \exp\Big(i \tan^{-1} \frac{\omega L}{R}\Big)$$

the current

$$I = \frac{V}{Z_{eq}} = \frac{V}{\sqrt{R^2 + (\omega L)^2}}\ \exp\Big(-i \tan^{-1} \frac{\omega L}{R}\Big) \tag{20.26}$$

and the voltages across the resistor and inductor

$$V_R = IZ_R = IR = V\,\frac{R}{\sqrt{R^2 + (\omega L)^2}}\ \exp\Big(-i \tan^{-1} \frac{\omega L}{R}\Big)$$

$$V_L = IZ_L = I(i\omega L) = V\,\frac{\omega L}{\sqrt{R^2 + (\omega L)^2}}\ \exp\Big(i \tan^{-1} \frac{R}{\omega L}\Big)$$

There are two things to note about the behavior of the AC LR circuit: First, whereas the current leads the voltage in a capacitive circuit, the current lags the voltage in an inductive circuit, as you can see from the opposite signs on the phases in eqq. (20.23) and (20.26): inductors don't like changes in current and retard them. Second, as you should be able to discern from the above relations, the behavior of the AC LR circuit as a function of driving angular frequency ω is entirely through the combination $\omega L/R$, which makes sense if we recall that the time constant of the DC LR circuit is L/R. For high frequencies ($\omega L/R \gg 1$), there isn't enough time to set a current flowing through the inductor and all of the applied voltage ends up backed up across the inductor. For low frequencies ($\omega L/R \ll 1$), there is more than enough time to change the flow of current, with the result that the applied voltage is entirely across the resistor, which alone limits the current flow.

20.2.3 The RLC Circuit

The analysis of the series RLC circuit is also virtually identical to that of the series RC circuit, except that there are now three contributions to the equivalent impedance and the calculations are considerably more tedious. You should find that the equivalent impedance is

$$\begin{aligned} Z_{eq} &= R + i\omega L + \frac{1}{i\omega C} \\ &= \sqrt{R^2 + \Big(\omega L - \frac{1}{\omega C}\Big)^2}\ \exp\Big(i \tan^{-1} \frac{1}{R}\Big(\omega L - \frac{1}{\omega C}\Big)\Big) \end{aligned}$$

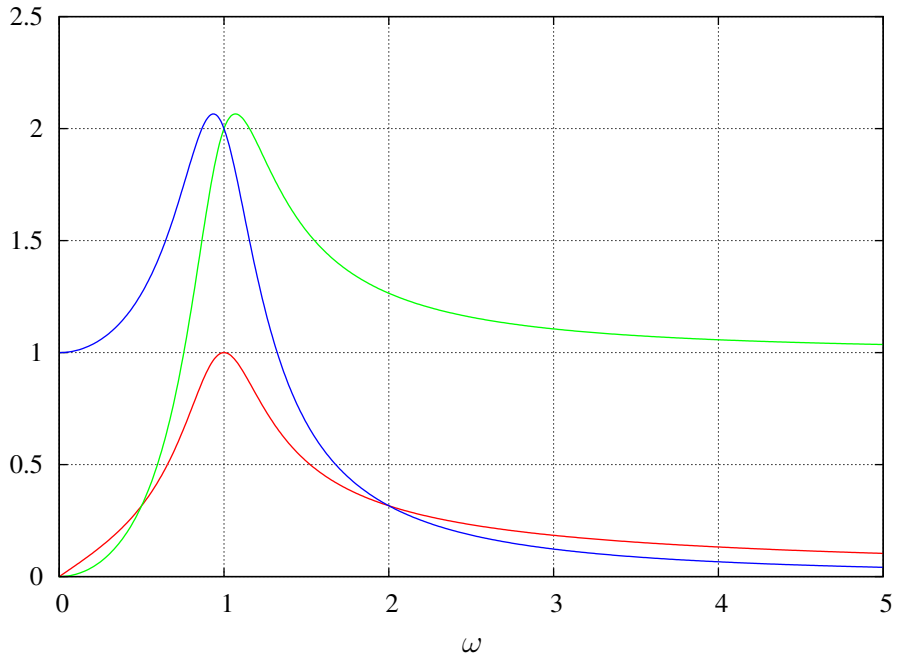


Figure 20.4: The Behavior of the RLC Circuit

the current

$$I = \frac{V}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}} \exp\left(-i \tan^{-1} \frac{1}{R}\left(\omega L - \frac{1}{\omega C}\right)\right)$$

and the voltages across the resistor, inductor, and capacitor

$$V_R = V \frac{R}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}} \exp\left(-i \tan^{-1} \frac{1}{R}\left(\omega L - \frac{1}{\omega C}\right)\right)$$

$$V_L = V \frac{\omega L}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}} \exp\left(i \tan^{-1} \frac{R}{\left(\omega L - \frac{1}{\omega C}\right)}\right)$$

$$V_C = V \frac{1}{\omega C \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}} \exp\left(-i \tan^{-1} \frac{R}{\left(\omega L - \frac{1}{\omega C}\right)}\right)$$

The ratios of the peak values of the voltages across the resistor, inductor, and capacitor to the peak value of the applied voltage are thus

$$\frac{V_{0R}}{V_0} = \frac{R}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}} \tag{20.27a}$$

$$\frac{V_{0L}}{V_0} = \frac{\omega L}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}} \tag{20.27b}$$

$$\frac{V_{0C}}{V_0} = \frac{1}{\omega C\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}} \tag{20.27c}$$

Plots of these ratios are shown, for some fake values of R, L, and C (and thus of units of ω on the horizontal axis) in fig. (20.4): V_{0R}/V_0 is the red curve, V_{0L}/V_0 green curve, and V_{0C}/V_0 the blue curve. For low driving frequencies ω, there is enough time for current not only to build but to fully charge the capacitor, with the result that the entire applied voltage ends up across the capacitor. At high frequencies, there isn't enough time to set a current flowing through the inductor and all of the applied voltage ends up backed up across the inductor. But more interesting than these extremes is the behavior near

$$\omega L - \frac{1}{\omega C} = 0$$

that is, for frequencies near

$$\omega = \frac{1}{\sqrt{LC}}$$

At this frequency the circuit has a resonance: there is a peak in the response of the circuit just like that of the damped, driven harmonic oscillator of §9.6.[7] This resonant behavior is useful for radio receivers:[8] The values of R, L, and C are adjusted to give a sharp peak at the frequency you want to receive. If the voltage across the resistor is the output, the components of the input signal matching this desired frequency are, as you can see from eq. (20.27a) or fig. (20.4), passed at full strength, while those of both higher and lower frequencies are strongly suppressed. Tuning the resonant frequency of the RLC circuit requires adjusting the value of either L or C, with the latter being the most practicable: when you are turning the tuning knob on the radio, you are changing the alignment of the plates in a variable capacitor — the more overlap between the plates, the larger the capacitance.

[7] We could, like most introductory texts, dwell on this similarity, showing how the equation governing the RLC circuit and its solution are identical in form to those for the damped, driven oscillator when certain analogies are made between R, L, and C and the various parameters of the oscillator. There is, however, little point to this exercise, so we won't go through it here.

[8] Or was, before everything went digital.

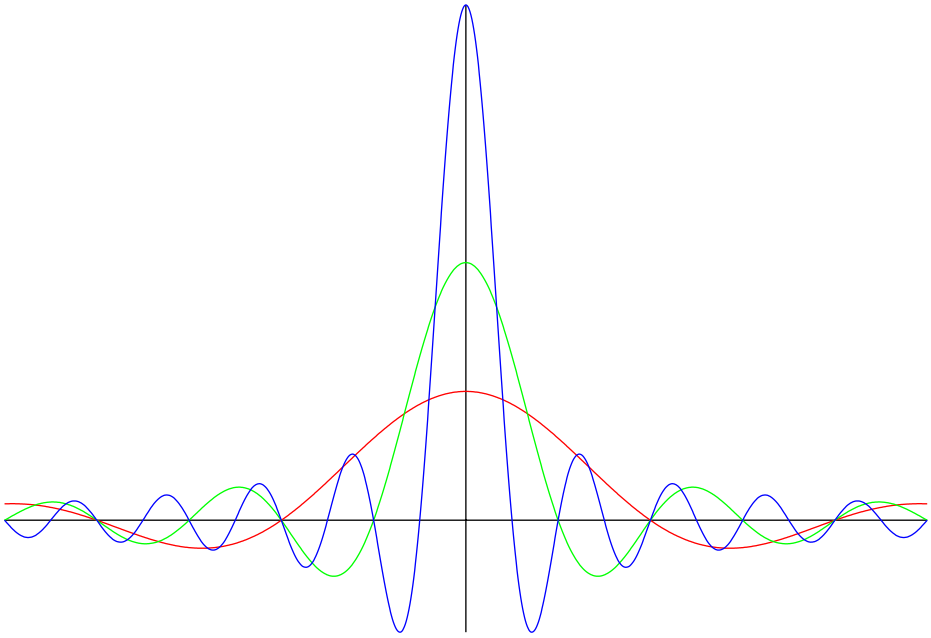


Figure 20.5: Portrait of an Aspiring Delta Function

20.3 The Sacred Rites of Initiation into the Awesome Mysteries of Delta Functions

We will now attempt to make the mysterious factor of 2π in

$$\int_{-\infty}^{\infty} d\omega\, e^{i\omega t} = 2\pi\delta(t) \tag{20.28}$$

at least plausible, though to call this a proof would make any self-respecting mathematician violently ill. Or perhaps just violent, depending on the mathematician.

Anyway, let us define an animal $\delta_n(t)$ by

$$\delta_n(t) = \frac{\sin nt}{\pi t}$$

Our claim is that

$$\lim_{n\to\infty} \delta_n(t) = \delta(t)$$

That is, that for any function $g(t)$

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} dt\, \delta_n(t)\, g(t) = g(0) \tag{20.29}$$

Fig. (20.5) shows a plot of our candidate $\delta_n(t)$ for three successive values of n, red being the lowest and blue the highest. As you can see from the plot,

$\delta_n(t)$ seems to have the right sort of behavior: as $n \to \infty$, the peak at $t = 0$ will become higher and narrower, and the wiggles to either side will become more rapid, with the result that the value of the rest of any integrand in which $\delta_n(t)$ is a factor will not vary significantly over any single oscillation of $\delta_n(t)$ and hence that the contribution of that oscillation to the integral will tend to cancel itself out. In other words, for large enough n the oscillations of $\sin nt$ between its peaks $+1$ and its troughs -1 will be so rapid for $t \neq 0$ that over each oscillation the variation in the function $g(t)$ of eq. (20.29) can be made arbitrarily small, so that the net contribution of

$$\delta_n(t)\, g(t)$$

over these oscillations will vanish: in the limit $n \to \infty$, the contribution from each oscillation will be

$$\frac{+1}{\pi t}\, g(t) + \frac{-1}{\pi t}\, g(t) = 0$$

In the limit $n \to \infty$ the only nonvanishing contribution to the integral of eq. (20.29) will therefore be from the central hump at $t = 0$.

This central hump occurs between the first zeros of $\sin(nt)$ on either side of $t = 0$, that is, at $nt = \pm\pi$ and hence $t = \pm\pi/n$. To see what happens at $t = 0$, consider

$$\begin{aligned}
\lim_{t\to 0} \delta_n(t) &= \lim_{t\to 0} \frac{\sin(nt)}{\pi t} \\
&= \lim_{t\to 0} \frac{nt - \frac{1}{3!}(nt)^3 + \frac{1}{5!}(nx)^5 - \cdots}{\pi t} \\
&= \frac{n}{\pi}
\end{aligned}$$

In the limit $n \to \infty$, the contribution of the central hump to the integral of eq. (20.29) will therefore effectively be that of a sawtooth, the height of which is n/π times $g(0)$ and the width of which is

$$\Delta x = \frac{\pi}{n} - \left(-\frac{\pi}{n}\right) = \frac{2\pi}{n}$$

This contribution is thus

$$(\text{area of sawtooth}) \times g(0) = \frac{1}{2}\left(\frac{2\pi}{n}\right)\left(\frac{n}{\pi}\right) g(0) = g(0) \quad \checkmark$$

We have thus established eq. (20.29). To connect this result with the integral of the complex exponential in eq. (20.28), the trick is to note that

$$\frac{1}{2\pi}\int_{-n}^{n} d\omega\, e^{i\omega t} = \frac{1}{2\pi}\int_{-n}^{n} d\omega\, (\cos\omega t + i \sin\omega t)$$

$$\begin{aligned}
&= \frac{1}{2\pi}\,\frac{1}{t}\Big[\sin\omega t - i\cos\omega t\Big]_{-n}^{n} \\
&= \frac{1}{2\pi}\,\frac{1}{t}\,2\sin nt \\
&= \frac{\sin nt}{\pi t} \\
&= \delta_n(t)
\end{aligned}$$

This gives us the desired result

$$\begin{aligned}
\delta(t) &= \lim_{n\to\infty}\delta_n(t) \\
&= \lim_{n\to\infty}\frac{1}{2\pi}\int_{-n}^{n} d\omega\, e^{i\omega t} \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega\, e^{i\omega t}
\end{aligned}$$

and hence

$$\int_{-\infty}^{\infty} d\omega\, e^{i\omega t} = 2\pi\delta(t)$$

Word.

Part V

And Now for Something Completely Different ...

Chapter 21

Lagrangian Dynamics

21.1 The Calculus of Variations

To derive the equations of motion in Lagrangian dynamics we will need the calculus of variations, which concerns itself with extremizing the values, not of functions of variables, but of functions of functions.[1] That may sound a bit obscure, but actually there aren't any concepts in the calculus of variations that can't quickly be developed from those already familiar to you from basic calculus and functions of variables, and we can in fact cover everything you need to know about the calculus of variations relatively painlessly in an example.

The variation δf in a function f is just an arbitrary infinitesimal change we make to f, and such variations behave just like differentials: if f is a function of a single variable x, then the variation δf in f due to a variation δx in x is

$$\delta f = \frac{df}{dx}\,\delta x$$

Similarly, if f is a function of two variables, x and y, the variation δf in f due to the variations δx and δy in x and y is

$$\delta f = \frac{\partial f}{\partial x}\,\delta x + \frac{\partial f}{\partial y}\,\delta y$$

Now, suppose we want to determine the shortest path in the xy plane from the origin to the point (a, b). Mathematically, this means determining the function $y(x)$ that minimizes the path length between the origin and (a, b).[2] Since the element ds of arc length in the xy plane is

$$ds = \sqrt{dx^2 + dy^2}$$

[1] Such functions of functions are known in the trade as functionals. Just in case you were curious.

[2] Here we see explicitly the difference between "ordinary" calculus, which can, by the usual sort of maxima-minima calculation, determine the value of the variable x that ex-

the distance s we travel along the path $y(x)$ will be given by

$$\begin{aligned} s &= \int_{(0,0)}^{(a,b)} ds \\ &= \int_{(0,0)}^{(a,b)} \sqrt{dx^2 + dy^2} \\ &= \int_{(0,0)}^{(a,b)} dx \sqrt{1 + \left(\frac{dy}{dx}\right)^2} \\ &= \int_{(0,0)}^{(a,b)} dx \sqrt{1 + y'^2} \end{aligned}$$

where we are using a prime to denote a derivative with respect to x.

How do we determine the $y(x)$ that minimizes the above integral? If we have some trial function $y(x)$, we can imagine generating a new trial function $y(x) + \delta y(x)$ by making an infinitesimal change $\delta y(x)$ to it. If $y(x)$ is *the* function that minimizes the path length, then *any* variation $\delta y(x)$ in $y(x)$ should result in a longer path length. In other words, the path length s is a minimum with respect to variations $\delta y(x)$ in $y(x)$. In "ordinary" calculus, if a function $f(x)$ is at a minimum then $df/dx = 0$ — which is equivalent to saying that the variation δf in f vanishes ($\delta f = 0$) for any variation δx in the variable x. So if our path length s is a minimum, then $\delta s = 0$ for all variations $\delta y(x)$ in $y(x)$:

$$0 = \delta s = \delta \int_{(0,0)}^{(a,b)} dx \sqrt{1 + y'^2} \tag{21.1}$$

To apply the δ to the integral, we note that the quantity we are varying is the function y (and hence also its derivatives y', y'', ...), so that the δ will be acting on the $\sqrt{1 + y'^2}$:

$$\delta \int_{(0,0)}^{(a,b)} dx \sqrt{1 + y'^2} = \int_{(0,0)}^{(a,b)} dx \, \delta\Big(\sqrt{1 + y'^2}\Big)$$

Since δ is just like a differential, the chain rule gives us

$$\int_{(0,0)}^{(a,b)} dx \, \delta\Big(\sqrt{1 + y'^2}\Big) = \int_{(0,0)}^{(a,b)} dx \, \frac{\partial\Big(\sqrt{1 + y'^2}\Big)}{\partial y'} \, \delta(y')$$

so that eq. (21.1) becomes

$$0 = \int_{(0,0)}^{(a,b)} dx \, \frac{\partial\Big(\sqrt{1 + y'^2}\Big)}{\partial y'} \, \delta(y') \tag{21.2}$$

tremizes a (known) function $y(x)$, and the calculus of variations, which can determine the *function* $y(x)$ itself that extremizes some quantity (in this case, that minimizes the distance from the origin to the point (a, b)).

Eq. (21.2) is telling us that the integral on the right-hand side vanishes *for all possible* variations $\delta(y')$ in $y'(x)$, and this can only be true if the factor

$$\frac{\partial\left(\sqrt{1+y'^2}\right)}{\partial y'}$$

in the integrand vanishes: if this factor did not vanish, we could find some $\delta(y')$ for which the integral would be nonzero. We therefore conclude that

$$\frac{\partial\left(\sqrt{1+y'^2}\right)}{\partial y'} = 0$$

which in turn means that

$$\sqrt{1+y'^2} = \text{const}$$

and hence that

$$y' = \text{const}$$

Thus y must be of the form $y = c_1 x + c_2$, where c_1 and c_2 are constants. We could impose the condition that $y = c_1 x + c_2$ pass through the origin and the point (a, b) to obtain $c_1 = b/a$ and $c_2 = 0$, but the result $y = c_1 x + c_2$ is already enough to establish that the shortest distance between two points is a straight line.

As, however, an illustration of some further techniques in the calculus of variations that will shortly prove useful, let us return to eq. (21.2) and do a little more gerrymandering. First, the variation of the derivative should be the same as the derivative of the variation, so that we can interchange the order of variations and derivatives: [3]

$$\delta(y') = \delta\left(\frac{dy}{dx}\right) = \frac{d}{dx}(\delta y)$$

Thus

$$dx\,\delta(y') = dx\,\frac{d}{dx}(\delta y) = d(\delta y)$$

and eq. (21.2) becomes

$$0 = \int_{(0,0)}^{(a,b)} \frac{\partial\left(\sqrt{1+y'^2}\right)}{\partial y'}\,d(\delta y)$$

[3] If this seems vaguely fishy or fishily vague to you, you can write the variation δy in the form $\epsilon\eta$, where η is an arbitrary function and ϵ is an infinitesimal constant. The variation δy then takes $y(x) \to y(x) + \epsilon\eta(x)$. If we apply d/dx to this, we get $y'(x) \to y'(x) + \epsilon\eta'(x)$, so that the variation $\delta(y')$ in the derivative of y is $\epsilon\eta'(x) = (\delta y)'$.

To tame the $d(\delta y)$, we can do an integration by parts:

$$\int u\,dv = uv - \int v\,du$$

With

$$u = \frac{\partial\left(\sqrt{1+y'^2}\right)}{\partial y'} \qquad\qquad v = \delta y$$

we can rewrite eq. (21.2) as

$$0 = \left[\frac{\partial\left(\sqrt{1+y'^2}\right)}{\partial y'}\,\delta y\right]_{(0,0)}^{(a,b)} - \int_{(0,0)}^{(a,b)} d\left(\frac{\partial\left(\sqrt{1+y'^2}\right)}{\partial y'}\right)\delta y \tag{21.3}$$

Note now that the first term on the right-hand side — the boundary term that is evaluated at the endpoints $(0,0)$ and (a,b) — has a factor of δy, and δy must vanish at these endpoints: our variations $\delta y(x)$ in the path $y(x)$ between $(0,0)$ and (a,b) are otherwise arbitrary, but they must not cause a shift in those endpoints — the whole issue is to determine the shortest path between those points, so allowing them to vary would be nonsense. The first term on the right-hand side therefore vanishes. And the remaining term we can rewrite as

$$0 = -\int_{(0,0)}^{(a,b)} d\left(\frac{\partial\left(\sqrt{1+y'^2}\right)}{\partial y'}\right)\delta y = -\int_{(0,0)}^{(a,b)} dx\,\frac{d}{dx}\left(\frac{\partial\left(\sqrt{1+y'^2}\right)}{\partial y'}\right)\delta y$$

From here, our reasoning is the same as before: since the integral on the right-hand side must vanish for all possible variations δy, we must have

$$\frac{d}{dx}\left(\frac{\partial\left(\sqrt{1+y'^2}\right)}{\partial y'}\right) = 0$$

which means that

$$\frac{\partial\left(\sqrt{1+y'^2}\right)}{\partial y'} = \text{const} \tag{21.4}$$

We could now do out the derivative and attempt to solve the resulting equation for $y(x)$, but this equation would turn out to be a fairly hairy nonlinear differential equation. There is, however, no need to use brute force and get all sweaty; we can already see from eq. (21.4) what happens: integrating eq. (21.4) will yield an algebraic equation for y' involving only constants, so the solution must be of the form

$$y' = \text{const}$$

just as we had obtained above.

21.2 The Brachistochrone

We work through the calculation for the brachistochrone because it is such a classic application of the calculus of variations, but the sections devoted to it are entirely an aside; you don't need anything in them for Lagrangian dynamics.

Anyway, as the name implies, the brachistochrone is the path of least time — specifically the curve (that is, the shape of "hill") down which the time it takes an object to slide under the influence of the gravitational force mg is minimal.

We set up our curve so that it starts at the origin of an xy plane and then descends to the point $(a, -b)$, where a and b are both positive. Since the distance through which the object has fallen at any given point along the curve is $h = -y$, by conservation of energy the speed of the object is $v = \sqrt{2gh} = \sqrt{-2gy}$. Using this, together with the same expression for the element ds of arc length as in the preceding section, we obtain for the time T of descent

$$\begin{aligned} T &= \int_{\text{arc}} dt \\ &= \int_{\text{arc}} \frac{ds}{v} \\ &= \int_{\text{arc}} \frac{\sqrt{dx^2 + dy^2}}{\sqrt{-2gy}} \\ &= \int_{x=0}^{a} \frac{\sqrt{1 + \left(\frac{dy}{dx}\right)^2}\, dx}{\sqrt{-2gy}} \\ &= \frac{1}{\sqrt{2g}} \int_0^a dx \, \sqrt{-\frac{1 + y'^2}{y}} \\ &= \frac{1}{\sqrt{2g}} \int_0^a dx \, \varphi \end{aligned}$$

where in the last line we have defined — just for convenience, so that we don't have to write it out all the time —

$$\varphi = \sqrt{-\frac{1 + y'^2}{y}}$$

Minimizing T by setting its variation to zero thus yields

$$0 = \delta T = \delta\left(\frac{1}{\sqrt{2g}} \int_0^a dx \, \varphi\right) = \frac{1}{\sqrt{2g}} \int_0^a dx \, \delta\varphi \tag{21.5}$$

To deal with the $\delta\varphi$, we note that $\varphi = \varphi(y, y')$, so that

$$\delta\varphi = \frac{\partial\varphi}{\partial y}\,\delta y + \frac{\partial\varphi}{\partial y'}\,\delta(y')$$

which, if we again use

$$\delta(y') = \delta\left(\frac{dy}{dx}\right) = \frac{d}{dx}(\delta y)$$

becomes

$$\delta\varphi = \frac{\partial\varphi}{\partial y}\,\delta y + \frac{\partial\varphi}{\partial y'}\,\frac{d}{dx}(\delta y)$$

Using this in eq. (21.5), we have

$$0 = \frac{1}{\sqrt{2g}}\int_0^a dx\left(\frac{\partial\varphi}{\partial y}\,\delta y + \frac{\partial\varphi}{\partial y'}\,\frac{d}{dx}(\delta y)\right) \tag{21.6}$$

To deal with the $d(\delta y)/dx$, we again integrate by parts:

$$\int_0^a dx\,\frac{\partial\varphi}{\partial y'}\,\frac{d}{dx}(\delta y) = \int_0^a \frac{\partial\varphi}{\partial y'}\,d(\delta y) = \left[\frac{\partial\varphi}{\partial y'}\,\delta y\right]_0^a - \int_0^a d\left(\frac{\partial\varphi}{\partial y'}\right)\delta y$$

Since our endpoints are fixed, our variation δy in the shape of the hill must vanish at $x = 0$ and $x = a$. The first term on the right-hand side therefore vanishes. And if we rewrite the second term as

$$-\int_0^a d\left(\frac{\partial\varphi}{\partial y'}\right)\delta y = -\int_0^a dx\,\frac{d}{dx}\left(\frac{\partial\varphi}{\partial y'}\right)\delta y$$

then we have

$$\int_0^a dx\,\frac{\partial\varphi}{\partial y'}\,\frac{d}{dx}(\delta y) = -\int_0^a dx\,\frac{d}{dx}\left(\frac{\partial\varphi}{\partial y'}\right)\delta y$$

and hence eq. (21.6) becomes

$$0 = \frac{1}{\sqrt{2g}}\int_0^a dx\left(\frac{\partial\varphi}{\partial y} - \frac{d}{dx}\left(\frac{\partial\varphi}{\partial y'}\right)\right)\delta y$$

Since this integration must vanish for all possible variations δy, we must have

$$\frac{\partial\varphi}{\partial y} - \frac{d}{dx}\left(\frac{\partial\varphi}{\partial y'}\right) = 0 \tag{21.7}$$

We could now plug in our definition

$$\varphi = \sqrt{-\frac{1 + y'^2}{y}}$$

into eq. (21.7) and solve the resulting equation by brute force — the resulting differential equation for y, while fairly hairy and nonlinear, would not really be any worse than the equation we will obtain below — , but there is a clever trick we can use to rewrite eq. (21.7) as a total derivative, and as this trick has more general application beyond the special case of the brachistochrone, we will invoke it here. Consider

$$y' \frac{\partial \varphi}{\partial y'} - \varphi \tag{21.8}$$

If we use the product rule and also note that, since $\varphi = \varphi(y, y')$,

$$\frac{d\varphi}{dx} = \frac{\partial \varphi}{\partial y}\frac{dy}{dx} + \frac{\partial \varphi}{\partial y'}\frac{dy'}{dx} = \frac{\partial \varphi}{\partial y}\, y' + \frac{\partial \varphi}{\partial y'}\, y''$$

then we obtain, when we expand out the derivative of (21.8),

$$\begin{aligned}\frac{d}{dx}\left(y' \frac{\partial \varphi}{\partial y'} - \varphi\right) &= \frac{dy'}{dx}\frac{\partial \varphi}{\partial y'} + y' \frac{d}{dx}\left(\frac{\partial \varphi}{\partial y'}\right) - \frac{d\varphi}{dx} \\ &= y'' \frac{\partial \varphi}{\partial y'} + y' \frac{d}{dx}\left(\frac{\partial \varphi}{\partial y'}\right) - \left(\frac{\partial \varphi}{\partial y}\, y' + \frac{\partial \varphi}{\partial y'}\, y''\right)\end{aligned}$$

The first and last terms on the right-hand side cancel, so that, pulling the common factor of y' out of the remaining terms, we are left with

$$\frac{d}{dx}\left(y' \frac{\partial \varphi}{\partial y'} - \varphi\right) = y'\left(\frac{d}{dx}\left(\frac{\partial \varphi}{\partial y'}\right) - \frac{\partial \varphi}{\partial y}\right)$$

On the right-hand side, the factor in parentheses is identical to the left-hand side of eq. (21.7), which means that

$$\frac{d}{dx}\left(y' \frac{\partial \varphi}{\partial y'} - \varphi\right) = 0$$

and hence that

$$y' \frac{\partial \varphi}{\partial y'} - \varphi = C \tag{21.9}$$

where C is a constant.

To solve eq. (21.9), we now revert to

$$\varphi = \sqrt{-\frac{1 + y'^2}{y}}$$

and do out the derivative by the chain rule:

$$\begin{aligned}C &= y' \frac{\partial \varphi}{\partial y'} - \varphi \\ &= y' \frac{\partial}{\partial y'}\left(\sqrt{-\frac{1 + y'^2}{y}}\right) - \sqrt{-\frac{1 + y'^2}{y}}\end{aligned}$$

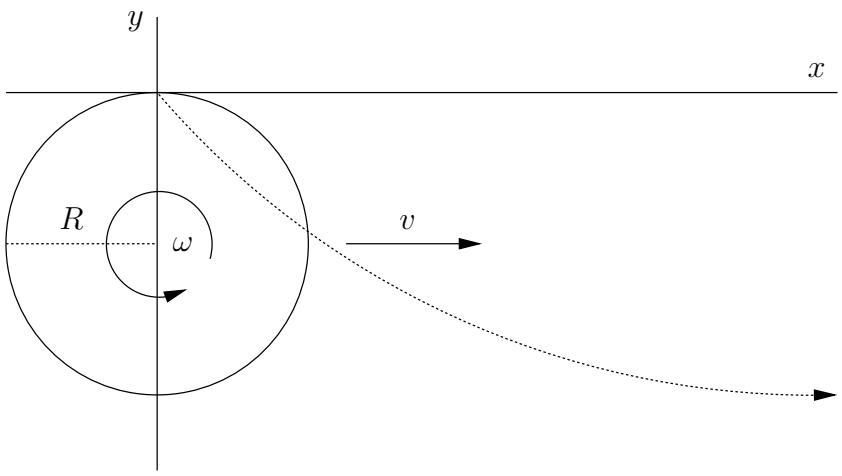


Figure 21.1: Tracing out a Cycloid

$$= y' \frac{1}{\sqrt{-y}} \frac{\partial}{\partial y'} \left(\sqrt{1+y'^2} \right) - \sqrt{-\frac{1+y'^2}{y}}$$
$$= y' \frac{1}{\sqrt{-y}} \frac{2y'}{2\sqrt{1+y'^2}} - \sqrt{-\frac{1+y'^2}{y}}$$

which, if we pull out a common factor of $1/\sqrt{-y(1+y'^2)}$, reduces to

$$= \frac{1}{\sqrt{-y(1+y'^2)}} \Big(y'^2 - (1+y'^2) \Big)$$
$$= -\frac{1}{\sqrt{-y(1+y'^2)}}$$

or

$$y(1+y'^2) = -\frac{1}{C^2} \tag{21.10}$$

As we will now verify, the solution to eq. (21.10) turns out to be a cycloid, the shape traced out by a point on the rim of a rolling wheel. Fig. (21.1) shows a wheel of radius R rolling underneath the x axis at a constant angular speed ω and hence a constant translational speed $v = \omega R$. The trajectory of the point that starts at the origin is shown by the dotted curve in fig. (21.1). If we parametrize the trajectory by $\theta = \omega t$, the coordinates of this point will be

$$\begin{aligned} x &= vt - R\sin\theta & \qquad y &= R\cos\theta - R \\ &= R(\omega t - \sin\omega t) & &= R(\cos\omega t - 1) \end{aligned} \tag{21.11}$$

To see this, note that the center of the oscillatory y motion is $y = -R$ and that the x coordinate must share in the forward motion of the whole wheel and hence translate forward a distance vt. All that remains is to work out the sin and cos parts of x and y: we know that x and y each have to have a term like either $\pm\sin\theta$ or $\pm\cos\theta$, and we can determine which sign and trig function goes with which coordinate by matching up the values and behavior we know x and y must at $\theta = 0$. In order to start at $y = 0$ at $\theta = 0$, we must have a $+R\cos\theta$ contribution to y. Similarly, in order to start at $x = 0$ at $\theta = 0$, we must have a $\pm R\sin\theta$ contribution to x, and of these two possibilities we must choose $-R\sin\theta$ in order for the forward translational velocity v to be canceled by the backward rotational velocity ωR when we evaluate dx/dt at $t = 0$.

The cycloid relations (21.11) give us

$$y' = \frac{dy}{dx} = \frac{dy/dt}{dx/dt} = \frac{-\omega R\sin\omega t}{\omega R(1-\cos\omega t)} = \frac{-\sin\omega t}{1-\cos\omega t}$$

so that

$$\begin{aligned}
y(1+y'^2) &= R(\cos\omega t - 1)\left(1 + \left(\frac{-\sin\omega t}{1-\cos\omega t}\right)^2\right) \\
&= R(\cos\omega t - 1)\left(1 + \frac{\sin^2\omega t}{(1-\cos\omega t)^2}\right) \\
&= R(\cos\omega t - 1)\frac{(1-\cos\omega t)^2 + \sin^2\omega t}{(1-\cos\omega t)^2} \\
&= \frac{R}{\cos\omega t - 1}\Big((1 - 2\cos\omega t + \cos^2\omega t) + \sin^2\omega t\Big) \\
&= \frac{R}{\cos\omega t - 1}\Big(1 - 2\cos\omega t + (\cos^2\omega t + \sin^2\omega t)\Big) \\
&= \frac{R}{\cos\omega t - 1}(1 - 2\cos\omega t + 1) \\
&= \frac{R}{\cos\omega t - 1}\,2(1-\cos\omega t) \\
&= -2R
\end{aligned}$$

Inasmuch as the right-hand side is a negative constant, the cycloid is indeed a solution of the brachistochrone eq. (21.10) and therefore is the shape that minimizes the time of descent. Just in case it comes up in a game of Trivial Pursuit.

21.2.1 A Brief Digression, For Those Inclined to Digress

The brachistochrone also has the interesting property that no matter where the object starts its slide, it will reach the bottom in the same amount of time. The brachistochrone is thus also known as the tautochrone (meaning *same time*) or, sometimes, the isochrone (meaning *equal time*), although this latter term is also used to mean a different sort of curve.[4]

To establish this tautochrone property, we return to the time relation

$$T = \int_{\text{arc}} dt = \int_{\text{arc}} \frac{ds}{v}$$

For the purpose of explicitly carrying out this integration, it will be convenient to express the cycloid curve in terms of a new variable $\tau = \omega t$, so that our previous relations

$$x = R(\omega t - \sin \omega t) \qquad\qquad y = R(\cos \omega t - 1)$$

for the curve become

$$x = R(\tau - \sin \tau) \qquad\qquad y = R(\cos \tau - 1)$$

Each point on the cycloid then has a corresponding value of τ; the section of the cycloid we are concerned with runs from $\tau = 0$ to $\tau = \pi$, corresponding to the extreme points $(0, 0)$ and $(\pi R, -2R)$, which are, respectively, the highest and lowest points on the cycloid curve. Suppose now that we are starting from rest from some arbitrary intermediate point on the cycloid curve that has coordinates

$$x_0 = R(\tau_0 - \sin \tau_0) \qquad\qquad y_0 = R(\cos \tau_0 - 1)$$

corresponding to some τ_0 that lies in the range $0 \leq \tau_0 \leq \pi$. The vertical distance that the object has descended when it has reached the point (x, y) corresponding to τ is then

$$\begin{aligned} y_0 - y &= R(\cos \tau_0 - 1) - R(\cos \tau - 1) \\ &= R(\cos \tau_0 - \cos \tau) \end{aligned}$$

Its velocity at this point is therefore

$$v = \sqrt{2g(y_0 - y)} = \sqrt{2gR(\cos \tau_0 - \cos \tau)}$$

[4] Namely, the curve along which the object will, in equal intervals of time, descend equal vertical distances.

In terms of τ, the element of arc length is

$$\begin{aligned} ds &= \sqrt{dx^2 + dy^2} \\ &= d\tau \sqrt{\left(\frac{dx}{d\tau}\right)^2 + \left(\frac{dy}{d\tau}\right)^2} \\ &= d\tau \sqrt{\Big(R(1-\cos\tau)\Big)^2 + (-R\sin\tau)^2} \\ &= d\tau\, R\sqrt{(1-\cos\tau)^2 + \sin^2\tau} \\ &= d\tau\, R\sqrt{1 - 2\cos\tau + \cos^2\tau + \sin^2\tau} \\ &= d\tau\, R\sqrt{1 - 2\cos\tau + 1} \\ &= d\tau\, R\sqrt{2(1-\cos\tau)} \end{aligned}$$

Putting our results for the time, velocity, and element of arc length together, we then have for the time T to go from location τ_0 to the bottom (at $\tau = \pi$)

$$\begin{aligned} T &= \int_{\tau=\tau_0}^{\pi} \frac{ds}{v} \\ &= \int_{\tau_0}^{\pi} d\tau\, \frac{R\sqrt{2(1-\cos\tau)}}{\sqrt{2gR(\cos\tau_0 - \cos\tau)}} \\ &= \sqrt{\frac{R}{g}} \int_{\tau_0}^{\pi} d\tau \sqrt{\frac{1-\cos\tau}{\cos\tau_0 - \cos\tau}} \end{aligned}$$

This integration is not only doable, but not even all that painful — if, with a little prescient finesse, we note that the double-angle trigonometric relations allow us to write

$$\begin{aligned} 1 - \cos\tau &= 1 - \cos\Big(2\cdot\tfrac{1}{2}\tau\Big) \\ &= 1 - \Big(1 - 2\sin^2\tfrac{1}{2}\tau\Big) \\ &= 2\sin^2\tfrac{1}{2}\tau \end{aligned}$$

and

$$\begin{aligned} \cos\tau_0 - \cos\tau &= \cos\Big(2\cdot\tfrac{1}{2}\tau_0\Big) - \cos\Big(2\cdot\tfrac{1}{2}\tau\Big) \\ &= \Big(2\cos^2\tfrac{1}{2}\tau_0 - 1\Big) - \Big(2\cos^2\tfrac{1}{2}\tau - 1\Big) \\ &= 2\Big(\cos^2\tfrac{1}{2}\tau_0 - \cos^2\tfrac{1}{2}\tau\Big) \end{aligned}$$

Our integral for the time T thus becomes

$$T = \sqrt{\frac{R}{g}} \int_{\tau_0}^{\pi} d\tau \sqrt{\frac{2\sin^2\frac{1}{2}\tau}{2\Big(\cos^2\frac{1}{2}\tau_0 - \cos^2\frac{1}{2}\tau\Big)}} = \sqrt{\frac{R}{g}} \int_{\tau_0}^{\pi} d\tau\, \frac{\sin\frac{1}{2}\tau}{\sqrt{\cos^2\frac{1}{2}\tau_0 - \cos^2\frac{1}{2}\tau}}$$

This may not look much more palatable, but hold my beer and watch this: if we make the change in variables

$$u = \frac{\cos\frac{1}{2}\tau}{\cos\frac{1}{2}\tau_0} \qquad\qquad u\Big|_{\tau=\tau_0} = \frac{\cos\frac{1}{2}\tau_0}{\cos\frac{1}{2}\tau_0} = 1$$

$$du = -\frac{1}{2}\,\frac{\sin\frac{1}{2}\tau}{\cos\frac{1}{2}\tau_0}\,d\tau \qquad\qquad u\Big|_{\tau=\pi} = -\frac{\cos\frac{\pi}{2}}{\cos\frac{1}{2}\tau_0} = 0$$

then from the relations for u and du we have

$$\cos\tfrac{1}{2}\tau = u\cos\tfrac{1}{2}\tau_0 \qquad\qquad \sin\tfrac{1}{2}\tau\;d\tau = -2\cos\tfrac{1}{2}\tau_0\;du$$

and our integral for the time T further reduces to

$$\begin{aligned}
T &= \sqrt{\frac{R}{g}}\int_{\tau_0}^{\pi}\sin\tfrac{1}{2}\tau\,d\tau\,\frac{1}{\sqrt{\cos^2\frac{1}{2}\tau_0 - \cos^2\frac{1}{2}\tau}} \\
&= \sqrt{\frac{R}{g}}\int_1^0 -2\cos\tfrac{1}{2}\tau_0\;du\,\frac{1}{\sqrt{\cos^2\frac{1}{2}\tau_0 - u^2\cos^2\frac{1}{2}\tau_0}} \\
&= -2\sqrt{\frac{R}{g}}\int_1^0 du\,\frac{1}{\sqrt{1-u^2}} \\
&= 2\sqrt{\frac{R}{g}}\int_0^1 du\,\frac{1}{\sqrt{1-u^2}}
\end{aligned}$$

where in the last step we have eaten the overall negative sign by reversing the limits. The integral we are now left with is simply that for the arcsin:

$$T = 2\sqrt{\frac{R}{g}}\;\sin^{-1}u\Big|_0^1 = 2\sqrt{\frac{R}{g}}\left(\frac{\pi}{2} - 0\right) = \pi\sqrt{\frac{R}{g}}$$

Since this is independent of the location τ_0 (that is, of the point (x_0, y_0) on the cycloid) from which the object started out, the time to reach the bottom of the cycloid is the same no matter how high up the cycloid the object is released. You can verify this on the wooden cycloid we have in the classroom by simultaneously releasing two balls from rest at different points on the cycloid and seeing that they do indeed both reach the bottom of the cycloid at the same time.

Thus endeth today's digression.

21.3 Lagrangian Dynamics

Recall that the total energy E of a system is the sum of its kinetic energy K and potential energy U:

$$E = K + U$$

What was significant about the total energy E was that it was conserved, that is, constant in value.

Though it might seem like an obscurely odd thing to do, we will define the Lagrangian L for a system to be the difference between the kinetic and potential energies:

$$L = T - U$$

where we have switched our notation for the kinetic energy from K to the T more conventionally used in the context of Lagrangian dynamics. This Lagrangian is always regarded as a function of the system's coordinates and the corresponding velocities. Thus for an object on an x axis, the Lagrangian would be regarded as a function of the object's coordinate x and the corresponding velocity $v = dx/dt = \dot{x}$. If, for example, we were dealing with a mass m attached to a spring of spring constant k, the Lagrangian would be

$$L = T - U = \tfrac{1}{2}m\dot{x}^2 - \tfrac{1}{2}kx^2$$

More generally, the coordinates of the system are conventionally denoted by q_i, $i = 1, 2, 3, \ldots$, and the corresponding velocities by $\dot{q}_i = dq_i/dt$. The q_i and $\dot{q}_i$ are called the generalized coordinates and generalized velocities. The action S is defined to be the time integral of the Lagrangian:

$$S = \int dt\ L(q_i, \dot{q}_i)$$

The principle of least action states that the evolution (that is, the motion, the trajectory) of the system is such that the action S is minimal. This principle, in conjunction with the above definitions of the Lagrangian and action, will turn out not only to reproduce Newton's laws of motion but to be much more general and powerful than Newton's laws; all of physics — mechanics, electromagnetism, quantum field theory, string theory, etc. — can be formulated by the principle of least action. It's the big-people way of doing physics.

Our first task will be to work out the consequences of the principle of least action by setting the variation δS to zero to minimize the action:

$$0 = \delta S = \delta \int_{t_1}^{t_2} dt\ L(q_i, \dot{q}_i)$$

The limits t_1 and t_2 on the integration are any two fixed times — that is, we can choose t_1 and t_2 to be any two times that we want, but we of course then

have to keep them fixed at these values when we do our variational calculation to determine the trajectory of the system between these two times. For simplicity, we will suppose that the system has only one coordinate q; later we will return to the more general case. For the single-coordinate case, we have, using the same techniques that we did in the previous sections of this chapter,

$$\begin{aligned}
0 &= \delta \int_{t_1}^{t_2} dt\ L(q, \dot{q}) \\
&= \int_{t_1}^{t_2} dt\ \delta L(q, \dot{q}) \\
&= \int_{t_1}^{t_2} dt \left(\frac{\partial L}{\partial q}\, \delta q + \frac{\partial L}{\partial \dot{q}}\, \delta \dot{q} \right) \\
&= \int_{t_1}^{t_2} dt \left(\frac{\partial L}{\partial q}\, \delta q + \frac{\partial L}{\partial \dot{q}}\, \delta \left(\frac{dq}{dt} \right) \right) \\
&= \int_{t_1}^{t_2} dt \left(\frac{\partial L}{\partial q}\, \delta q + \frac{\partial L}{\partial \dot{q}}\, \frac{d}{dt}(\delta q) \right) \\
&= \int_{t_1}^{t_2} dt\ \frac{\partial L}{\partial q}\, \delta q + \int_{t=t_1}^{t_2} \frac{\partial L}{\partial \dot{q}}\, d(\delta q)
\end{aligned} \tag{21.12}$$

To deal with the second term, we do the usual integration by parts:

$$\int u\, dv = uv - \int v\, du$$

With

$$u = \frac{\partial L}{\partial \dot{q}} \qquad\qquad v = \delta q$$

we have

$$du = d\left(\frac{\partial L}{\partial \dot{q}} \right) = dt\, \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} \right)$$

so that

$$\int_{t=t_1}^{t_2} \frac{\partial L}{\partial \dot{q}}\, d(\delta q) = \left[\frac{\partial L}{\partial \dot{q}}\, \delta q \right]_{t=t_1}^{t_2} - \int_{t_1}^{t_2} dt\ \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} \right)\, \delta q$$

We now need to remember that we are keeping are endpoints fixed, so that $\delta q = 0$ at $t = t_1$ and $t = t_2$: what we are trying to determine is the trajectory of the system in between these endpoints.[5] This kills the first term, so that we are left with

$$\int_{t=t_1}^{t_2} \frac{\partial L}{\partial \dot{q}}\, d(\delta q) = - \int_{t_1}^{t_2} dt\ \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}} \right)\, \delta q$$

[5] You might question how we can keep the coordinates of the endpoints fixed when we don't yet know the trajectory of the system: although we could certainly posit some initial location for the system at $t = t_1$, how would we know the final location at $t = t_2$?

Using this in eq. (21.12) gives

$$\begin{aligned} 0 &= \int_{t_1}^{t_2} dt\, \frac{\partial L}{\partial q}\,\delta q - \int_{t_1}^{t_2} dt\, \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right)\,\delta q \\ &= \int_{t_1}^{t_2} dt\, \left(\frac{\partial L}{\partial q} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right)\right)\,\delta q \end{aligned}$$

In order for this integral to vanish for all variations δq in the trajectory, we must have

$$0 = \frac{\partial L}{\partial q} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right)$$

or, as it is more usually written,

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = 0 \tag{21.13}$$

Eq. (21.13) is the equation that governs the evolution of the system; it is variously called the equation of motion, the Euler-Lagrange equation,[6] or simply the Lagrange equation. When the system has more than one coordinate q_i, there will, since the variations δq_i in these coordinates are independent of each other, be one such equation for each coordinate:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0 \qquad\qquad (i = 1,\, 2,\, 3,\, \ldots)$$

In the sense that each coordinate represents a "direction" along which the system can move, the coordinates q_i are sometimes referred to as degrees of freedom.

Before we do out a couple of examples to show how Lagrangian dynamics is applied and how it reproduces Newton's laws, just a couple of other quick points:

First, the quantity $\partial L/\partial q_i$ is, for reasons that will become clear when we do out the examples, defined to be the conjugate momentum p_i corresponding to the coordinate q_i (or, less circumlocutiously, the momentum conjugate to q_i):

$$p_i = \frac{\partial L}{\partial \dot{q}_i}$$

The answer is that we don't have to: we are simply supposing that the final location is in fact on the trajectory of the system — for our purposes, it is only important that that final location is fixed; we don't have to know exactly where it is. The principle of least action will turn out to give us a differential equation for the motion of the system, so that, starting from its initial location at $t = t_1$, we can trace out the trajectory one infinitesimal step at a time and thus figure out where the system will in fact be at $t = t_2$.

[6] In spite of the similarity in spelling between Leonard *Eu*ler and Ferris B*ue*ller, *Euler* is pronounced like *oiler* — do not say "Yew-ler" unless you want to sound like a redneck.

Second, coordinates on which the Lagrangian does not explicitly depend are known as cyclic coordinates. Since $\partial L/\partial q_i = 0$ for such coordinates, the Lagrange equation corresponding to them becomes

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) = \frac{dp_i}{dt}$$

which means that the momentum p_i conjugate to a cyclic coordinate q_i is conserved.

In its root sense, the term *cyclic coordinate* refers to an angular coordinate, which is literally cyclic in the sense of being periodic. Often one is concerned with the motion of a mass around a center of attraction — a comet passing by the Sun or an electron orbiting around a nucleus — , and in such cases the force exerted on the mass depends only on the radial distance between the mass and the center of attraction and not on any angle of orientation. Such forces are known as central forces. The potential energies corresponding to central forces are, like the forces themselves, functions only of the radial variable r, so that the resulting Lagrangians are independent of angular variables like the spherical coordinates θ and φ, and the momenta conjugate to these angular variables are therefore conserved. As will be seen in the second example below, the momentum conjugate to an angular variable turns out, as you might expect, to be an angular momentum. The cyclicity of angular variables and the conservation of the angular momenta conjugate to them corresponds physically to the absence of torque for a central force: since the force $\mathbf{F}$ is along the radial direction, the torque $\boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}$ vanishes, which in turn, since $\boldsymbol{\tau} = d\mathbf{L}/dt$, means that the angular momentum $\mathbf{L}$ is constant.

The term *cyclic coordinate* has, however, come to be applied much more generally than just to angular coordinates in central-force problems; by analogy, any coordinate variable on which the Lagrangian does not explicitly depend is called cyclic, whether that variable is periodic or not. One could, we suppose, get all righteously indignant about this apparent abuse of the word *cyclic*, but there is no need: those with delicate sensibilities will be happy to learn that the equivalent term ignorable coordinate is also in general use. Whatever you choose, after careful moral reflection, to call such coordinates, don't forget that the essential point is that the momentum conjugate to any coordinate on which the Lagrangian does not explicitly depend is conserved.

And now, the moment you've all been waiting for ...

Suppose, for example, we have a mass m attached to a spring of spring constant k. As noted above, the Lagrangian for this system is

$$L = T - U = \tfrac{1}{2}m\dot{x}^2 - \tfrac{1}{2}kx^2$$

Since

$$\frac{\partial L}{\partial x} = \frac{\partial}{\partial x}\left(\tfrac{1}{2}m\dot{x}^2 - \tfrac{1}{2}kx^2\right) = -kx$$
$$\frac{\partial L}{\partial \dot{x}} = \frac{\partial}{\partial \dot{x}}\left(\tfrac{1}{2}m\dot{x}^2 - \tfrac{1}{2}kx^2\right) = m\dot{x}$$

the Lagrange equation for this system is

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = 0$$
$$\frac{d}{dt}(m\dot{x}) - (-kx) = 0$$
$$m\ddot{x} + kx = 0$$

or

$$-kx = m\frac{d^2x}{dt^2}$$

which is just $F = ma$ with F being the spring force $-kx$. And the momentum conjugate to x is, from the above,

$$p = \frac{\partial L}{\partial \dot{x}} = m\dot{x}$$

which is just the familiar $p = mv$.

As a second example, consider a mass m moving under the influence of the gravitational force mg. If we set things up in cylindrical coordinates (r, θ, z) with the positive z axis oriented vertically upward, then the potential energy of the mass is mgz and its squared velocity is [7]

$$\begin{aligned} v^2 &= \mathbf{v}\cdot\mathbf{v} \\ &= \left(\frac{dr}{dt}\,\widehat{\mathbf{r}} + r\frac{d\theta}{dt}\,\widehat{\boldsymbol{\theta}} + \frac{dz}{dt}\,\widehat{\mathbf{z}}\right)^2 \\ &= \left(\frac{dr}{dt}\right)^2 + r^2\left(\frac{d\theta}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 \\ &= \dot{r}^2 + r^2\dot{\theta}^2 + \dot{z}^2 \end{aligned}$$

The Lagrangian for this system is thus

$$L = T - U = \tfrac{1}{2}mv^2 - mgz = \tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2 + \dot{z}^2) - mgz$$

[7] If you have forgotten how the r and θ components of the velocity work out, see eq. (3.41) on p.151.

There will be a Lagrange equation for each of the three coordinates r, θ, and z. For the z coordinate we have

$$\frac{\partial L}{\partial z} = \frac{\partial}{\partial z}\Big(\tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2 + \dot{z}^2) - mgz\Big) = -mg$$
$$\frac{\partial L}{\partial \dot{z}} = \frac{\partial}{\partial \dot{z}}\Big(\tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2 + \dot{z}^2) - mgz\Big) = m\dot{z}$$

so that the corresponding Lagrange equation is

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{z}}\right) - \frac{\partial L}{\partial z} = 0$$
$$\frac{d}{dt}(m\dot{z}) - (-mg) = 0$$
$$m\ddot{z} + mg = 0$$

or $\ddot{z} = -g$, exactly as we would have expected: the gravitational force mg produces a downward acceleration of magnitude g.

Likewise for the radial coordinate r we have

$$\frac{\partial L}{\partial r} = \frac{\partial}{\partial r}\Big(\tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2 + \dot{z}^2) - mgz\Big) = mr\dot{\theta}^2$$
$$\frac{\partial L}{\partial \dot{r}} = \frac{\partial}{\partial \dot{r}}\Big(\tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2 + \dot{z}^2) - mgz\Big) = m\dot{r}$$

so that the corresponding Lagrange equation is

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{r}}\right) - \frac{\partial L}{\partial r} = 0$$
$$\frac{d}{dt}(m\dot{r}) - mr\dot{\theta}^2 = 0$$
$$m\ddot{r} - mr\dot{\theta}^2 = 0$$

or

$$mr\dot{\theta}^2 = m\ddot{r}$$

If we recall that $\omega = \dot{\theta}$, we see that $mr\dot{\theta}^2 = m\omega^2 r$ is just the familiar centrifugal force. There being no physical forces with a radial component, this is the only contribution to $m\ddot{r}$.

Unlike r and z, θ does not occur explicitly in the Lagrangian, that is, θ is a cyclic coordinate. The momentum p_θ conjugate to θ is therefore conserved:

$$p_\theta = \frac{\partial L}{\partial \dot{\theta}} = \frac{\partial}{\partial \dot{\theta}}\Big(\tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2 + \dot{z}^2) - mgz\Big) = mr^2\dot{\theta} = \text{const}$$

If we note that mr^2 is the point mass's moment of inertia I and that $\dot{\theta} = \omega$, we see that p_θ is indeed just the familiar angular momentum $L = I\omega$.

As noted above, Lagrangian dynamics does not simply reproduce Newton's laws — Lagrangian dynamics is far more general and powerful than Newton's laws: not only classical Newtonian mechanics, but also special and general relativity, the Maxwell equations and electromagnetism, quantum mechanics, quantum field theories, and string theories can be formulated in Lagrangian terms. In fact, the more fundamental and closer a theory is to the truth — that is, to being the ultimate physical theory — , the more naturally it can be formulated in terms of the principle of least action. The ultimate object of physics is to write down the Lagrangian that accounts for all of the physics in our universe. It has been proved mathematically that for each symmetry in the Lagrangian, there is a corresponding conserved quantity — in Appendix B, for example, we show that the invariance of physics under spacetime translations corresponds to conservation of energy and momentum. The quest to discover the Lagrangian that governs the universe is thus a quest to discover the fundamental symmetry that underlies nature. Under the assumption that anything allowed mathematically should occur in nature, one seeks to answer the question, What is the most general possible Lagrangian for the universe? Although it might seem that the possibilities are limitless, it has been shown that there are in fact only five possible string theories, and these five string theories have been shown to be just five different representations of a single underlying theory known as M-theory. It may well turn out that, far from there being limitless possibilities, our universe is the way it is because that's the only way it could be.

21.4 Lagrange Multipliers & Constraints

A discussion of the treatment of constraints in Lagrangian dynamics is arguably going a bit far afield, but Lagrange multipliers are kind of neat, so what the ****.

A generalized coordinate can also be referred to as a degree of freedom, in the sense that the coordinate represents a "direction" along which the system is free to move. There are, however, not always as many degrees of freedom as there are coordinates: sometimes the motion of the system is constrained, with each constraint corresponding to a relation among the coordinates that reduces the number of independent variables. The motion of a mass on a frictionless incline, for example, can be set up in terms of two coordinates, x for the horizontal direction and y for the vertical, but the motions in the x and y directions are not independent: the condition that the mass remain on the incline constitutes a constraint on the system, a constraint that can be expressed as $y/x = \tan\theta$ for an incline at angle θ to the horizontal. With this constraint between the two coordinates, there remains really only $2 - 1 = 1$ degree of freedom, and the Lagrangian could instead have been expressed in

terms of a single coordinate along an axis parallel to the incline.

Whenever possible, it is, of course, preferable to express the Lagrangian in terms of coordinates that are independent of each other. But there are occasions when there is no simple set of fully independent coordinates, and in those cases one instead expresses the Lagrangian in terms of coordinates that are not fully independent but related by one or more constraints. The method of Lagrange multipliers gives us a way of taking these constraints into account when we generate the Lagrange equations for the system.

As an example to illustrate the method, suppose we want to determine the shortest distance from the origin to the curve $y = x^2 + 1$. If you consider the plot of $y = x^2 + 1$, it is obvious that the shortest distance is along the vertical segment that goes straight up the y axis from $(0, 0)$ to $(0, 1)$. But we can also arrive at this result by a powerful technique of very general application: What we want is to minimize the distance $\sqrt{x^2 + y^2}$ — or equivalently the squared distance $x^2 + y^2$ — from the origin to a point (x, y), but only for points (x, y) that lie on the curve $y = x^2 + 1$. One way to impose this constraint is to first write it in the form $\phi(x, y) = 0$ and then add to the function being extremized a term $\lambda\phi(x, y)$, treating λ, which is called the Lagrange multiplier, as an additional independent variable. In this case, we are trying to minimize $x^2 + y^2$ subject to the constraint $y = x^2 + 1$, so we first write the constraint as $y - x^2 - 1 = 0$ and instead minimize the function

$$f(x, y, \lambda) = x^2 + y^2 + \lambda(y - x^2 - 1)$$

Doing the usual maxima-minima thing on this function of three variables, we have

$$\begin{aligned} 0 &= \frac{\partial f}{\partial x} = 2x + \lambda(-2x) = 2x(\lambda + 1) \\ 0 &= \frac{\partial f}{\partial y} = 2y + \lambda \\ 0 &= \frac{\partial f}{\partial \lambda} = y - x^2 - 1 \end{aligned}$$

As you can see, extremizing with respect to λ always gives us back the original constraint equation, thus guaranteeing that the maxima-minima solution obeys the constraint. Solving this set of equations for x, y, and λ yields three sets of values, two of which have to be rejected because they involve imaginary x. The remaining set of values is $x = 0$, $y = 1$, $\lambda = -2$, corresponding to the solution $(x, y) = (0, 1)$ that we expected.

This is the method of Lagrange multipliers: you express the constraint in the form $\phi(x, y) = 0$ and add $\lambda\phi(x, y)$ to the function that you are extremizing, treating the λ as an additional independent variable. If you have n constraints, you express them as n relations of the form $\phi_i(x, y) = 0$,

$i = 1, 2, 3, \ldots, n$, and add n terms of the form $\lambda_i \phi_i(x, y)$ to the function that you are extremizing, treating the λ_i as n additional independent variables.

This same technique can be used in Lagrangian dynamics for dealing with systems subject to constraints: when minimizing the action to generate the Lagrange equations for a system with coordinates q_i, we express each constraint in the form $\phi(q_i) = 0$ and then generate the Lagrange equations from the modified Lagrangian

$$L = T - U + \lambda\phi$$

treating λ as an additional independent coordinate that therefore has its own Lagrange equation:

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\lambda}}\right) - \frac{\partial L}{\partial \lambda} = -\phi$$

The Lagrange equation corresponding to the "coordinate" λ thus guarantees that the constraint $\phi = 0$ will be obeyed.

Consider, for example, a mass m on a frictionless incline at angle θ to the horizontal. Expressed in terms of the horizontal coordinate x and vertical coordinate y, the Lagrangian is

$$L = \tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2) - mgy$$

If we put our origin on the incline, the constraint that the mass moves along the incline is $y/x = \tan\theta$, which we will write in the form [8]

$$\phi(x, y) = x\sin\theta - y\cos\theta = 0$$

For the modified Lagrangian

$$L = \tfrac{1}{2}m(\dot{x}^2 + \dot{y}^2) - mgy + \lambda(x\sin\theta - y\cos\theta)$$

the Lagrange equations are then

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = m\ddot{x} - \lambda\sin\theta \tag{21.14a}$$

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{y}}\right) - \frac{\partial L}{\partial y} = m\ddot{y} + mg + \lambda\cos\theta \tag{21.14b}$$

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\lambda}}\right) - \frac{\partial L}{\partial \lambda} = -x\sin\theta + y\cos\theta \tag{21.14c}$$

[8] Why this form and not $y - x\tan\theta = 0$ or any of the infinitely many other equivalent expressions? We could use other such expressions, but you will see at the end of the calculation that there is a point to using the one we have chosen. As is so often the case in life, you are being cynically manipulated.

This constitutes a system of three coupled differential equations for x, y, and λ. We could go about solving these equations many ways: we could note that the second time derivative of eq. (21.14c) gives

$$0 = -\ddot{x}\sin\theta + \ddot{y}\cos\theta$$

and hence $\ddot{x} = \ddot{y}\cot\theta$. Using this in eq. (21.14a), we obtain

$$0 = m\ddot{y}\cot\theta - \lambda\sin\theta$$

which yields

$$\lambda = m\ddot{y}\,\frac{\cot\theta}{\sin\theta} = m\ddot{y}\,\frac{\cos\theta}{\sin^2\theta}$$

Using this result for λ in eq. (21.14b), we then have

$$\begin{aligned} 0 &= m\ddot{y} + mg + \lambda\cos\theta \\ &= m\ddot{y} + mg + m\ddot{y}\,\frac{\cos\theta}{\sin^2\theta}\cos\theta \\ &= m\ddot{y}\,\frac{\sin^2\theta + \cos^2\theta}{\sin^2\theta} + mg \\ &= m\ddot{y}\csc^2\theta + mg \end{aligned}$$

Solving this for $\ddot{y}$, we finally arrive at

$$\ddot{y} = -g\sin^2\theta$$

and, by back-substitution,

$$\begin{aligned} \ddot{x} &= \ddot{y}\cot\theta = -g\sin\theta\cos\theta \\ \lambda &= m\ddot{y}\,\frac{\cos\theta}{\sin^2\theta} = -mg\cos\theta \end{aligned}$$

Though they may look a little strange, these solutions for $\ddot{x}$ and $\ddot{y}$ are just what we expected: the horizontal and vertical components of an acceleration $g\sin\theta$ down the incline are $-g\sin\theta\cos\theta$ and $-g\sin\theta\sin\theta = -g\sin^2\theta$, respectively. And it is no accident that the result for λ looks suspiciously like the normal force $mg\cos\theta$:[9] as we will see in the next section, the value of a Lagrange multiplier, in conjunction with the expression ϕ for the constraint, gives us a result for the force responsible for making the system obey the constraint. In this example, it is the normal force that prevents the mass from sinking down into the incline and ensures that the motion is along the incline.

[9] Had we added a term $-\lambda\phi$ to the Lagrangian rather than $+\lambda\phi$, we would have obtained $\lambda = mg\cos\theta$ rather than $\lambda = -mg\cos\theta$. But then our sign convention for the constraint term added to the Lagrangian would have differed from that of most of the rest of the world, and the expression that we derive in the next section for forces of constraint would have a rather unæsthetic negative sign. So as it turns out you weren't actually being all that cynically manipulated after all.

21.5 Forces of Constraint

We know that the Lagrange-multiplier term $\lambda\phi$ modifies the Lagrangian in a way that fully takes into account the constraint; our task now is to discover the relation between this $\lambda\phi$ term and the force of constraint that literally forces the system to obey the constraint. Because the Lagrangian has dimensions of energy, we expect that the $\lambda\phi$ term will be related to the work done by the force of constraint. To investigate this further, let us look at the variation of the modified Lagrangian $L = T - U + \lambda\phi$:

$$\begin{aligned}\delta L &= \delta T - \delta U + \delta(\lambda\phi)\\ &= \delta T - \delta U + \delta\lambda\,\phi + \lambda\,\delta\phi\end{aligned}$$

Since the Lagrangian is a function of the physical generalized coordinates q_i, $i = 1, 2, 3, \ldots, n$, the corresponding generalized velocities $\dot{q}_i$, and the artificially introduced unphysical coordinate λ, the various contributions to this δL will ultimately depend on the variations δq_i in the physical coordinates and the variation $\delta\lambda$ in the unphysical coordinate λ. The $\delta\lambda\,\phi$ term is the sole contribution to δL from $\delta\lambda$ and corresponds to the Lagrange equation that gives $\phi = 0$; it has nothing to do with the work done by the force of constraint through the displacements δq_i of the physical coordinates and is therefore not the contribution in which we are interested. That leaves the $\lambda\,\delta\phi$ term. Since ϕ is a function of the coordinates q_i, by the chain rule the variation in ϕ is

$$\delta\phi = \sum_{i=1}^{n} \frac{\partial\phi}{\partial q_i}\,\delta q_i$$

so that the contribution of the $\lambda\,\delta\phi$ term to δL is

$$+\lambda\,\delta\phi = +\sum_{i=1}^{n} \lambda\frac{\partial\phi}{\partial q_i}\,\delta q_i \tag{21.15}$$

To see the relation between (21.15) and force, consider now the contribution that would be made to the $-\delta U$ term of δL by a force $\mathbf{F}$. Recall that for a displacement $d\mathbf{r}$ the contributions dU to the potential energy associated with a force $\mathbf{F}$ and dW to the work done by the force are

$$dU = -\mathbf{F}\cdot d\mathbf{r} \qquad\qquad dW = \mathbf{F}\cdot d\mathbf{r} = -dU$$

Here we are working, not with displacements $d\mathbf{r}$ along the trajectory of the system, but with arbitrary variations δq_i in that trajectory, so that the analogue of the displacement $d\mathbf{r}$ is $\delta\mathbf{q} = (\delta q_1, \delta q_2, \delta q_3, \ldots)$. The corresponding contribution $-\delta U$ of the variation in the potential energy to δL is thus

$$-\delta U = +\sum_{i=1}^{n} F_i\,\delta q_i = +\delta W$$

where δW is the variation in the work done by force $\mathbf{F}$. The contribution to δL from the work done by the force of constraint should therefore be of the same form and sign:

$$+\delta W_{constraint} = +\sum_{i=1}^{n} F_{i,\,constraint}\,\delta q_i \tag{21.16}$$

Comparing (21.16) with (21.15), we see that $\lambda\,\delta\phi$ term in δL must therefore represent this work:

$$\delta W_{constraint} = \sum_{i=1}^{n} F_{i,\,constraint}\,\delta q_i = \sum_{i=1}^{n} \lambda \frac{\partial \phi}{\partial q_i}\,\delta q_i \tag{21.17}$$

Since eq. (21.17) must hold for all possible variations $\delta\mathbf{q} = (\delta q_1, \delta q_2, \delta q_3, \dots)$, we conclude that the components of the force of constraint (for which the conventional notation is, believe it or not, $\mathbf{Q}$) are given by

$$Q_i = \lambda \frac{\partial \phi}{\partial q_i} \tag{21.18}$$

In the example of the mass sliding down a frictionless incline in the preceding section, we had

$$\lambda = -mg\cos\theta \qquad\qquad \phi(x,y) = x\sin\theta - y\cos\theta$$

so that the x and y components of the force of constraint are

$$\begin{aligned} Q_x &= \lambda \frac{\partial}{\partial x}(x\sin\theta - y\cos\theta) \\ &= -mg\cos\theta\,\frac{\partial}{\partial x}(x\sin\theta - y\cos\theta) \\ &= -mg\cos\theta\sin\theta \\ Q_y &= \lambda \frac{\partial}{\partial y}(x\sin\theta - y\cos\theta) \\ &= -mg\cos\theta\,\frac{\partial}{\partial y}(x\sin\theta - y\cos\theta) \\ &= mg\cos^2\theta \end{aligned}$$

As you can see from fig. (21.2), these are exactly what we expect for the x and y components of the normal force.

The connecting of bodies by cords also constrains the motion of a system: the motions of bodies connected by a cord are not independent, so that there is a mathematical relation among their coordinates, with the corresponding force of constraint being the tension. Consider, for example, Atwood's machine, depicted in fig. (21.3) in an exceedingly boring manner that should

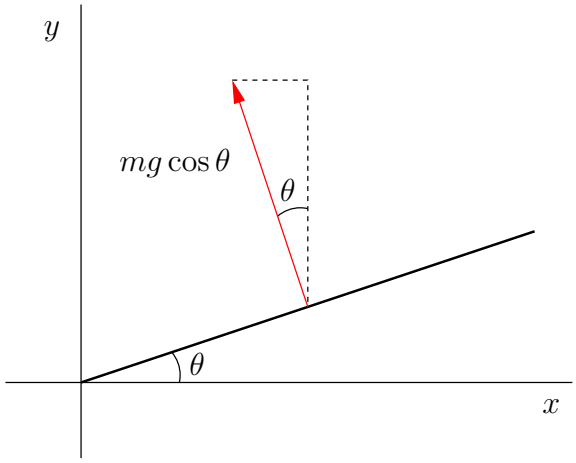


Figure 21.2: Revenge of the Incline

probably be a criminal offense. For simplicity, we will take the pulley to be massless and frictionless, so that there is a single tension T throughout the cord. We could, of course, take into account the relation between the motions of the two masses right from the start and set up the Lagrangian in terms of a single coordinate, but let us instead use distinct coordinates x_1 and x_2 for the heights of m_1 and m_2, as shown in fig. (21.3), with the positive x_1 and x_2 directions oriented vertically upward. The constraint that relates these coordinates is quite simple: we do not have to worry about the length of the cord or the level from which x_1 and x_2 are measured; it is enough to note that, the vertical displacements of m_1 and m_2 being equal and opposite, the sum of x_1 and x_2 is constant. If we denote this constant by ℓ, then our constraint is $x_1 + x_2 = \ell$ and we may take as our constraint function

$$\phi(x_1, x_2) = x_1 + x_2 - \ell = 0$$

Then our Lagrangian

$$\begin{aligned} L &= T - U + \lambda\phi \\ &= \tfrac{1}{2}m_1\dot{x}_1^2 + \tfrac{1}{2}m_2\dot{x}_2^2 - m_1 g x_1 - m_2 g x_2 + \lambda(x_1 + x_2 - \ell) \end{aligned}$$

yields the equations of motion

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_1}\right) - \frac{\partial L}{\partial x_1} = m_1\ddot{x}_1 + m_1 g - \lambda \tag{21.19a}$$

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}_2}\right) - \frac{\partial L}{\partial x_2} = m_2\ddot{x}_2 + m_2 g - \lambda \tag{21.19b}$$

$$0 = \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{\lambda}}\right) - \frac{\partial L}{\partial \lambda} = -(x_1 + x_2 - \ell) \tag{21.19c}$$

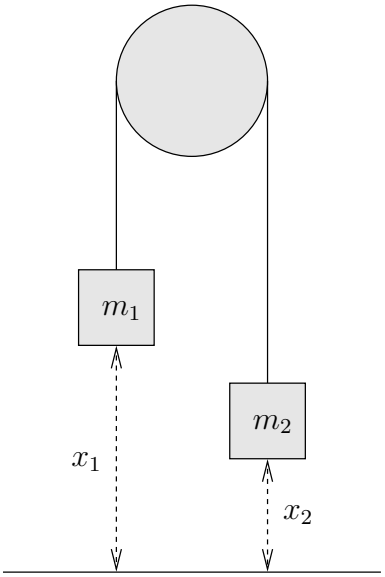


Figure 21.3: This is Really Getting Old

To solve this system of equations, we can take two time derivatives of eq. (21.19c) to obtain

$$0 = \ddot{x}_1 + \ddot{x}_2 \tag{21.20}$$

Subtracting eq. (21.19b) from eq. (21.19a) to eliminate λ and making use of eq. (21.20) in the form $\ddot{x}_2 = -\ddot{x}_1$ then gives

$$\begin{aligned} 0 &= m_1\ddot{x}_1 - m_2\ddot{x}_2 + (m_1 - m_2)g \\ &= (m_1 + m_2)\ddot{x}_1 + (m_1 - m_2)g \end{aligned}$$

so that we obtain the expected solutions for the accelerations of m_1 and m_2:

$$\ddot{x}_1 = -\frac{m_1 - m_2}{m_1 + m_2}\, g \qquad\qquad \ddot{x}_2 = -\ddot{x}_1 = \frac{m_1 - m_2}{m_1 + m_2}\, g$$

More interesting is the solution for λ: from eq. (21.19a) and our solution for $\ddot{x}_1$, we have

$$\lambda = m_1(\ddot{x}_1 + g) = m_1(-\frac{m_1 - m_2}{m_1 + m_2}\, g + g) = \frac{2m_1 m_2}{m_1 + m_2}\, g$$

By eq. (21.18), the x_1 component of the force of constraint is therefore

$$Q_{x_1} = \lambda \frac{\partial \phi}{\partial x_1} = \frac{2m_1 m_2}{m_1 + m_2}\, g\, \frac{\partial}{\partial x_1}(x_1 + x_2 - \ell) = \frac{2m_1 m_2}{m_1 + m_2}\, g$$

and similarly the x_2 component works out to

$$Q_{x_2} = \frac{2m_1 m_2}{m_1 + m_2}\, g$$

These components are, you will recall, the same as the result we had previously obtained for the tension by other methods. Since our convention was that positive x_1 and x_2 were upward, these positive components indicate that the tension is exerted in the upward direction on both m_1 and m_2.

21.6 Problems

1. A projectile of mass m moves solely under the influence of the gravitational force mg.

   (a) Set up the Lagrangian in terms of Cartesian coordinates, with a horizontal x axis and a vertical y axis.

   (b) Generate the Lagrange equation corresponding to the coordinate x.

   (c) Generate the Lagrange equation corresponding to the coordinate y.

   (d) What are the momenta conjugate to the coordinates x and y?

   (e) You should have found that x is a cyclic coordinate, so that its conjugate momentum is conserved. Make physical sense of this.

2. A mass m slides on a frictionless incline at angle θ to the horizontal. *Yawn.*

   (a) Since the motion is only along the incline, this system has only one degree of freedom, that is, can be described in terms of a single coordinate specifying your location along the incline. Express the Lagrangian of the system in terms of this coordinate.

   (b) Generate the corresponding Lagrange equation.

3. Recall that in polar coordinates (r, θ) the components of the velocity are

$$v_r = \frac{dr}{dt} = \dot{r} \qquad\qquad v_\theta = r\frac{d\theta}{dt} = r\dot{\theta}$$

   (a) Set up the Lagrangian for a satellite of mass m moving under the influence of the gravitational pull GMm/r^2 of a planet of mass $M \gg m$. You may for simplicity treat the mass M as stationary (or, if your conscience bothers you, set things up in terms of the reduced mass $\mu = mM/(m+M)$).

   (b) Generate the Lagrange equations corresponding to the coordinates r and θ.

   (c) Show that when the motion is circular the Lagrange equation for the radial coordinate reduces to what you would have gotten by setting up the usual circular-motion force relation. See the footnote if you need a hint.[10]

[10] For circular motion, the radius is constant. *Doh!*

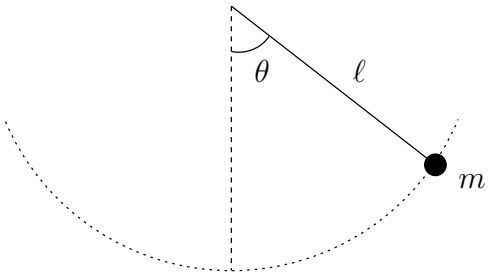


Figure 21.4: Problem 4

4. Fig. (21.4) shows a simple pendulum consisting of a point mass m on the end of a massless string of length ℓ.

   Determine the Lagrangian of the system, using θ as the sole coordinate, and generate the corresponding Lagrange equation.

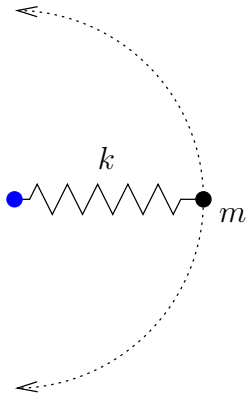


Figure 21.5: Problem 5

5. Fig. (21.5) shows a mass m attached to a spring of spring constant k and relaxed length ℓ that, in addition to stretching and compressing, can also swivel freely in the plane of the page about its other end, which is anchored at the blue point. Only the spring force acts on the mass.

   (a) Set up the Lagrangian for the system in polar coordinates.

   (b) Generate the Lagrange equations and interpret them physically. That is, to what physical force or effect does each of the terms correspond?

   (c) Suppose now that the gravitational force mg, directed vertically down the page, also acts on the mass. Determine the Lagrangian of the system and generate and interpret the Lagrange equations. For definiteness, take $\theta = 0$ to correspond to vertically downward.

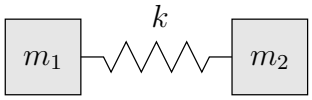


Figure 21.6: Problem 6

6. Fig. (21.6) shows two masses, on an x axis, connected by a spring. The spring has spring constant k and (rather unrealistically, but for simplicity) a relaxed length of zero. No other forces act on the masses.

   (a) Set up the Lagrangian of the system in terms of the individual x coordinates x_1 and x_2 of the masses.

   (b) Now set up the Lagrangian instead in terms the relative and center-of-mass coordinates

   $$x = x_2 - x_1 \qquad\qquad X = \frac{1}{m_1 + m_2}(m_1 x_1 + m_2 x_2)$$

   (If you are rusty on center-of-mass and relative coordinates, you might want to review eq. (6.22a) and friends on p.310.)

   (c) Generate the Lagrange equations corresponding to the coordinates x and X and interpret these equations of motion physically.

   (d) Suppose now that, instead of moving along an x axis, the two masses can move in the plane of the page.

      i. Set up the Lagrangian in terms of the Cartesian coordinates X and Y of the center of mass and the relative polar coordinates r and θ.

      ii. Generate the Lagrange equations and interpret these equations of motion physically.

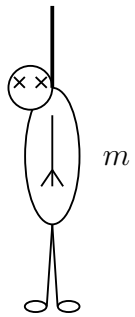


Figure 21.7: Problem 7

7. As a gentle warm-up exercise in constraints, consider a stationary mass m hanging from a cord, as shown in fig. (21.7).

   (a) Set up the modified Lagrangian $L = T - U + \lambda\phi$ in terms of the vertical coordinate y of the mass. If you are confused about what to use for the constraint function ϕ, see the footnote.[11]

   (b) Generate the Lagrange equations corresponding to y and the Lagrange multiplier λ.

   (c) Show that the force of constraint is, as expected, the tension $T = mg$, and that the sign for its direction is also as expected.

8. Continuing our gentle introduction to constraints, suppose a corpse of mass m moves in an xy plane under the influence of the gravitational force mg, with the positive y direction taken to be upward.

   (a) Set up the Lagrangian for the case that the corpse is constrained to move only along the horizontal line $y = a$, where a is a constant. Generate the equations of motion, solve them, and determine the force of constraint, including its direction.

   (b) Set up the Lagrangian for the case that the corpse is constrained to move only along the vertical line $x = a$, where a is a constant. Generate the equations of motion, solve them, and determine the force of constraint.

[11] The constraint relation in this case is so trivial that it might well confound you: $y = \text{const}$.

9. A mass m with no ambition or direction in life moves freely in a plane.

   (a) Set up the Lagrangian for this mass in polar coordinates (r, θ). (Note that the absence of a force on the mass means that $U = 0$.)

   (b) The mass now gets religion and, instead of wandering aimlessly, moves in a circle of radius R. Determine a constraint function ϕ corresponding to this constraint.

   (c) Generate the r, θ, and λ Lagrange equations for your modified Lagrangian $L = T - U + \lambda\phi$.

   (d) Interpret the Lagrange equation for θ in terms of angular momentum.

   (e) Determine the force of constraint and make physical sense of your result for it, including its sign for direction.

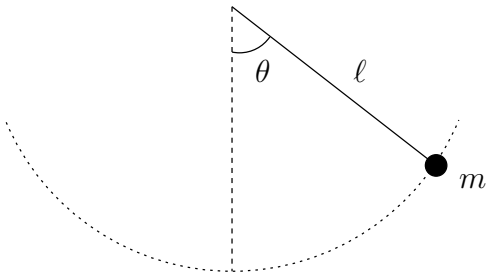


Figure 21.8: Problem 10

10. We will now revisit the simple pendulum, this time expressing the Lagrangian in terms of both of the polar coordinates r and θ and imposing the constraint $r = \ell$.

    (a) Set up the modified Lagrangian $L = T - U + \lambda\phi$ and generate the Lagrange equations for r, θ, and the Lagrange multiplier λ.

    (b) Interpret the Lagrange equation for θ in terms of gravitational torque and angular momentum, and show that it reproduces what you got in $^{\#}4$ when the constraint is imposed.

    (c) Solve for λ and show that it gives the expected result for the tension in the cord, including the sign for its direction.

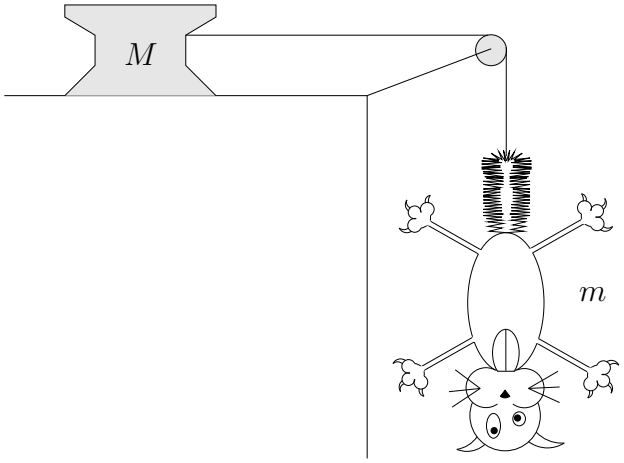


Figure 21.9: Problem 11

11. In another blast from even further in the past, we will now rework the cat-and-the-anvil problem of fig. (21.9) (#21 of p.229). Recall that the surface on which the anvil lies is level and frictionless and that the pulley is massless and frictionless.

    (a) Without worrying yet about the constraints, set up $T-U$ for the system in terms of the horizontal and vertical coordinates X and Y of the mass M and the vertical coordinate y of the mass m. Be sure to specify which are your positive X, Y, and y directions.

    (b) Now determine the constraints and include them in your Lagrangian. Note that there will be two constraints: one corresponding to the mass M being constrained to move on the surface, and one corresponding to the coupling of the motions of the two masses by the cord. Your Lagrangian will therefore be of the form $L = T - U + \lambda_1\phi_1 + \lambda_2\phi_2$. See the footnote if you need a hint.[12]

    (c) Generate the Lagrange equations for X, Y, y, λ_1, and λ_2.

    (d) Solve these equations to obtain results for $\ddot{X}$, $\ddot{y}$, and the Lagrange multipliers.

    (e) Show that the force of constraint corresponding to the mass M being constrained to move on the surface is the expected normal force and that it has the expected sign for its direction.

    (f) Show that the X and y components Q_X and Q_y of the force of constraint corresponding to the cord constraint are the expected tension. Show that the signs for the directions of Q_X and Q_y are also as expected.

[12] Depending on your sign conventions for directions, $X - y$ or the like is the length of the cord, which, while unknown, certainly doesn't change.

12. A uniform solid disk of mass m and radius R rolls without slipping down an incline at angle θ to the horizontal.

    (a) Without worrying yet about the constraint that the disk is rolling, set up $T - U$ for the disk in terms of the coordinates x and φ, where x is measured along the incline and φ is the angle through which the disk has rotated about its own axis as it rolls.

    (b) Determine a constraint relation ϕ between x and φ that corresponds to rolling without slipping and modify your Lagrangian accordingly.

    (c) Generate the Lagrange equations corresponding to x, φ, and the Lagrange multiplier and show that these yield the expected results for $\ddot{x}$ and $\ddot{\varphi}$.

    (d) Solve for the Lagrange multiplier and show that the components Q_x and Q_φ of the force of constraint are the expected frictional force and torque, with the expected signs for directions.

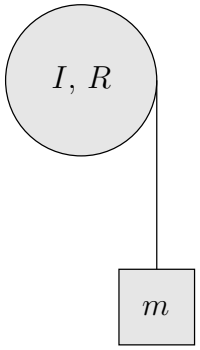


Figure 21.10: Problem 13

13. Fig. (21.10) shows an unapologetically boring mass m suspended from a cord wrapped around a frictionless and equally unimaginative pulley of radius R and moment of inertia I.

    (a) Set up the Lagrangian for this system, including the constraint that relates the angle through which the pulley has rotated to the distance that the mass has descended.

    (b) Generate the Lagrange equations and show that these yield the expected results for the acceleration of the mass m and the angular acceleration of the pulley.

    (c) Solve for the Lagrange multiplier and show that the components of the force of constraint are the expected tension and torque, with the expected signs for directions.

14. In this problem we will deal with a spherical pendulum, that is, a pendulum that can swing in three dimensions rather than just in a plane. Recall (eq. (2.4) on p.87) that the line element in spherical coordinates (r, θ, φ) is

$$d\mathbf{r} = dr\,\hat{\mathbf{r}} + r\,d\theta\,\widehat{\boldsymbol{\theta}} + r\sin\theta\,d\varphi\,\widehat{\boldsymbol{\varphi}}$$

Dividing this by dt and dotting it into itself, we therefore have for the squared velocity v^2 in spherical coordinates

$$v^2 = \left(\dot{r}\,\hat{\mathbf{r}} + r\dot{\theta}\,\widehat{\boldsymbol{\theta}} + r\sin\theta\,\dot{\varphi}\,\widehat{\boldsymbol{\varphi}}\right)^2 = \dot{r}^2 + r^2\dot{\theta}^2 + r^2\sin^2\theta\,\dot{\varphi}^2$$

For simplicity, we will take $\theta = 0$ to be oriented vertically downward. The length of the pendulum is ℓ and the bob is a point mass m.

(a) Set up the Lagrangian for the pendulum, including the constraint that $r = \ell$.

(b) Generate the Lagrange equations for r, θ, φ, and the Lagrange multiplier.

(c) Interpret the θ equation of motion in terms of gravitational torque and angular momentum.

(d) Interpret the φ equation of motion in terms of angular momentum.

(e) Determine the force of constraint and make physical sense of your result for it in terms of tension, weight, and centripetal effects.

15. A mass m slides under the influence of gravity down a frictionless helical track given in cylindrical coordinates by $(r, \theta, z) = (\rho, \theta, \kappa\theta)$, where ρ and κ are positive constants.

    (a) Set up the Lagrangian for the mass, including the constraint that the mass is on the helix. Note that constraining a mass moving in three dimensions to a trajectory along a one-dimensional curve means that you will need two constraints (both of which we essentially gave you just above). See the footnote if you need a hint about the kinetic term.[13]

    (b) Generate the Lagrange equations for r, θ, z, and the two Lagrange multipliers.

    (c) Solve these equations for $\ddot{z}$ and $\ddot{\theta}$.

    (d) Determine the components of the two forces of constraint and interpret these in terms of centripetal effects, normal forces, and torques.

    (e) Make physical sense of your results in the limits

        i. $\kappa \gg \rho$

        ii. $\kappa \ll \rho$

16. A mass m moves under the influence of gravity in an xy plane (with, as usual, x horizontal and y vertical) along a track, the shape of which is given by $y = f(x)$. (Although we have not specified f — it could be any old function of x — , it is to be treated as a given.)

    (a) Set up the Lagrangian, including the constraint that the mass follows the curve $y = f(x)$, and generate the equations of motion.

    (b) Relate $\ddot{y}$ and $\ddot{x}$ by taking time derivatives of your constraint relation.

    (c) Explain how you would go about solving for $\ddot{x}$, $\ddot{y}$, the Lagrange multiplier, and the components Q_x and Q_y of the force of constraint.

17. The motion of a mass m is confined to a frictionless surface defined by

$$z = f(x, y)$$

    where (x, y, z) are the Cartesian coordinates of the mass.

    Show that the force of constraint is always perpendicular to the surface. See the footnote if you need a hint.[14]

    Quite generally, the trajectory of a system is perpendicular to a constraint $\phi(q_i) = 0$, since such constraints are of the form $\phi(q_i) = \text{const}$.

[13] Remember that cylindrical coordinates are just polar coordinates plus a z coordinate: $v^2 = \dot{r}^2 + r^2\dot{\theta}^2 + \dot{z}^2$.

[14] You should find that the force of constraint is $\mathbf{F} = \lambda\nabla\phi$. Also recall that, by eq. (2.5) on p.94, for a displacement $d\mathbf{r}$, $d\phi = \nabla\phi \cdot d\mathbf{r}$.

18. (A preview of the Lagrangian formulation of field theory.)

The Lagrangian for a free (that is, noninteracting) relativistic scalar field φ of mass m is[15]

$$\mathcal{L} = \tfrac{1}{2}\partial^\mu\varphi\,\partial_\mu\varphi - \tfrac{1}{2}m^2\varphi^2$$

which is a concise four-vector notation for

$$\mathcal{L} = \tfrac{1}{2}\left(\left(\frac{\partial\varphi}{\partial t}\right)^2 - \left(\frac{\partial\varphi}{\partial x}\right)^2 - \left(\frac{\partial\varphi}{\partial y}\right)^2 - \left(\frac{\partial\varphi}{\partial z}\right)^2\right) - \tfrac{1}{2}m^2\varphi^2$$

When quantized, such a field would correspond to spinless (spin-0) particles of mass m, similar to the Higgs particle. The generalized coordinate in this Lagrangian is the field φ, not the spacetime coordinates x, y, z, and t, so that the Lagrange relation is

$$\partial^\mu\left(\frac{\partial\mathcal{L}}{\partial\partial^\mu\varphi}\right) - \frac{\partial\mathcal{L}}{\partial\varphi} = 0$$

which again is a concise four-vector notation for

$$\frac{\partial}{\partial t}\left(\frac{\partial\mathcal{L}}{\partial\left(\frac{\partial\varphi}{\partial t}\right)}\right) - \frac{\partial}{\partial x}\left(\frac{\partial\mathcal{L}}{\partial\left(\frac{\partial\varphi}{\partial x}\right)}\right) - \frac{\partial}{\partial y}\left(\frac{\partial\mathcal{L}}{\partial\left(\frac{\partial\varphi}{\partial y}\right)}\right) - \frac{\partial}{\partial z}\left(\frac{\partial\mathcal{L}}{\partial\left(\frac{\partial\varphi}{\partial z}\right)}\right) - \frac{\partial\mathcal{L}}{\partial\varphi} = 0$$

(a) Show that the Lagrange equation for the field φ is

$$\Box\varphi + m^2\varphi = 0$$

where the D'Alambertian $\Box$ is the spacetime equivalent of the three-dimensional operator ∇^2:

$$\nabla^2 = \nabla\cdot\nabla = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

$$\Box = \partial^\mu\partial_\mu = \frac{\partial^2}{\partial t^2} - \frac{\partial^2}{\partial x^2} - \frac{\partial^2}{\partial y^2} - \frac{\partial^2}{\partial z^2}$$

[15] Actually, this is the Lagrangian density, which is just the Lagrangian per unit volume. Since a field extends over all of space, to work with the values of the field at specific points in space we work with quantities like energy per unit volume, similar to the way we worked with charge density ρ and current density $\mathbf{j}$ when dealing with electric and magnetic fields in the Maxwell equations. But working per unit volume doesn't change the Lagrange relations. The fancy $\mathcal{L}$ that we have used is the conventional symbol for a Lagrangian density, to distinguish it from an ordinary Lagrangian L.

(b) A plane wave solution is a simple oscillatory solution of the form

$$\varphi = \varphi_0 \, e^{-ik_\mu x^\mu} = \varphi_0 \, e^{i(k_x x + k_y y + k_z z - \omega t)}$$

where φ_0 is the amplitude, ω the angular frequency, and

$$\mathbf{k} = k_x \, \hat{\mathbf{x}} + k_y \, \hat{\mathbf{y}} + k_z \, \hat{\mathbf{z}}$$

the wave number (the spatial equivalent of ω).

Show that the above plane wave satisfies the Lagrange equation as long as

$$\omega^2 - k^2 = m^2$$

According to quantum theory, the momentum $\mathbf{p}$ and energy E are related to the wave number $\mathbf{k}$ and angular frequency ω by $\mathbf{p} = \hbar\mathbf{k}$ and $E = \hbar\omega$, where $\hbar$ is Plank's constant divided by 2π. It is, however, customary to work in units where $\hbar = c = 1$, a system known as natural units, and we have in fact done this above. In natural units, the relations are simply $\mathbf{p} = \mathbf{k}$ and $E = \omega$. When the factors of the speed of light are put back in, $\omega^2 - k^2 = m^2$ is thus simply the familiar relativistic relation between energy, momentum, and mass:

$$E^2 - p^2c^2 = m^2c^4$$

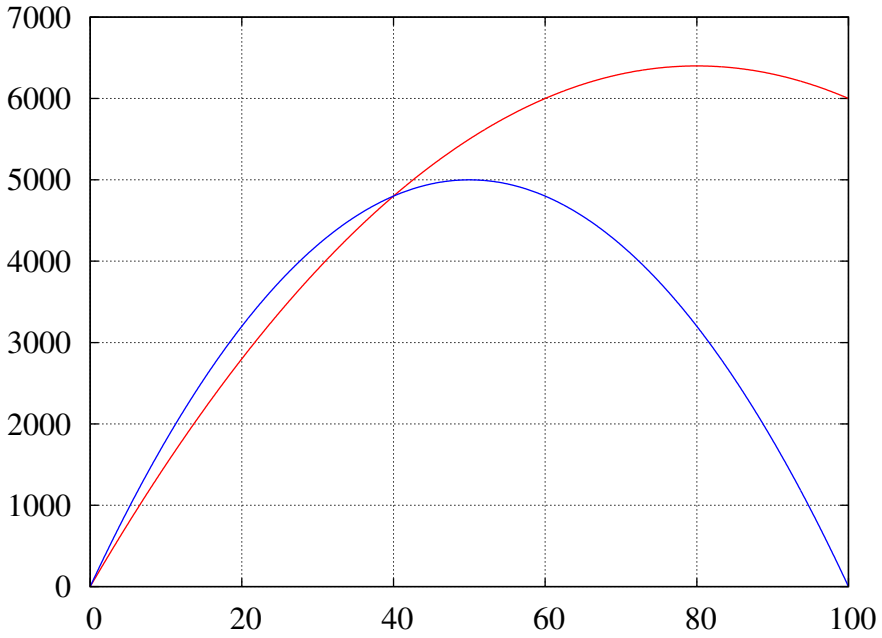


Figure 21.11: Problem 19

19. For the sake of an example of extremizing quantities subject to constraints from another discipline, let us turn for a moment to economics. Suppose that each machine in a factory can be tooled to produce either lawn gnomes or pink flamingos. Further suppose, rather artificially, that the production costs and demand curves for gnomes and flamingos are such that their profit functions are

$$p_g(n_g) = 160n_g - n_g^2 \qquad\qquad p_f(n_f) = 200n_f - 2n_f^2$$

where p_g and p_f are the dollars of profit made per day when n_g and n_f machines are producing gnomes and flamingos, respectively. Fig. (21.11) shows the plots of these profit functions, with gnomes in red and flamingos, rather perversely, in blue.

(a) Obviously in the best of all possible worlds, you would tool 80 machines to produce gnomes and 50 machines to produce flamingos. But suppose[16] that the factory has only 100 machines. Expressing this constraint in the form $\phi(n_g, n_f) = 0$, show that maximizing the total profit function

$$p(n_g, n_f) = 100n_g - n_g^2 + 240n_f - 3n_f^2 + \lambda\phi$$

yields $n_g = 60$, $n_f = 40$, and $\lambda = -40$.

[16] There has certainly been a lot of supposing going on around here. And note that this is just a maxima-minima problem with a constraint: there is no principle of least action in economics — nor any other discernible principle for that matter.

(b) If we formally differentiate the total profit function $p(n_g, n_f)$ with respect to ϕ, we see that

$$\frac{\partial p}{\partial \phi} = \lambda$$

is the marginal cost of the constraint ϕ. This rate of change of profit per unit constraint is known as the *shadow price* of the constraint. Show that at the point where profit is maximized,

$$\frac{\partial p_g}{\partial n_g} = \frac{\partial p_f}{\partial n_f} = \lambda = \frac{\partial p}{\partial \phi}$$

Make sense of this coincidence of rates and of the units of λ. What does λ mean in terms of dollars, machines, gnomes, and flamingos?

(c) Show that the same coincidence would occur and this same interpretation hold more generally for

$$p(n_g, n_f) = p_g(n_g) + p_f(n_f) + \lambda\phi$$

for any p_g and p_f.

21.7 Sketchy Answers

(2a) Modulo your sign convention for direction, $L = \frac{1}{2}m\dot{q}^2 - mgq\sin\theta$.

(4) $L = \frac{1}{2}m\ell^2\dot{\theta}^2 - mg\ell(1-\cos\theta)$.

(5a) $L = \frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) - \frac{1}{2}k(r-\ell)^2$.

(5c) $L = \frac{1}{2}m\dot{r}^2 + \frac{1}{2}mr^2\dot{\theta}^2 - \frac{1}{2}k(r-\ell)^2 + mgr\cos\theta$.

(6a) $L = \frac{1}{2}m_1\dot{x_1}^2 + \frac{1}{2}m_2\dot{x_2}^2 - \frac{1}{2}k(x_2-x_1)^2$.

(6b) $L = \frac{1}{2}\mu\dot{x}^2 + \frac{1}{2}M\dot{X}^2 - \frac{1}{2}kx^2$.

(6(d)i) $L = \frac{1}{2}\mu(\dot{r}^2 + r^2\dot{\theta}^2) + \frac{1}{2}M(\dot{X}^2 + \dot{Y}^2) - \frac{1}{2}kr^2$.

(11d) Modulo your sign conventions,

$$\ddot{X} = -\frac{m}{m+M}\,g = \ddot{y} \qquad \lambda_1 = Mg \qquad \lambda_2 = -\frac{mM}{m+M}\,g$$

(11f) With the previous sign conventions, $Q_X = -\dfrac{mM}{m+M}\,g = -Q_y$.

(12a) Modulo your sign convention for direction,

$$T - U = \tfrac{1}{2}m\dot{x}^2 + \tfrac{1}{4}mR^2\dot{\varphi}^2 + mgx\sin\theta$$

(12c) $\ddot{x} = \frac{2}{3}g\sin\theta, \quad \ddot{\varphi} = \frac{2}{3}\dfrac{g}{R}\sin\theta$.

(12d) $Q_x = -\frac{1}{3}mg\sin\theta, \quad Q_\varphi = \frac{1}{3}mgR\sin\theta$.

(13c) Modulo your sign conventions for directions, your components of the force of constraint should work out to

$$-\,mg\,\frac{I/mR^2}{1+I/mR^2} \qquad\qquad mgR\,\frac{I/mR^2}{1+I/mR^2}$$

(15c) $\ddot{z} = -\dfrac{\kappa^2}{\kappa^2+\rho^2}\,g, \quad \ddot{\theta} = -\dfrac{\kappa}{\kappa^2+\rho^2}\,g$.

(15d) Between the two forces of constraint, you should obtain

$$Q_r = -m\rho\dot{\theta}^2 \qquad Q_z = \frac{\rho^2}{\kappa^2+\rho^2}\,mg \qquad Q_\theta = -\frac{\kappa\rho^2}{\kappa^2+\rho^2}\,mg$$

Chapter 22

Real Physics, Unification Theories, & Cosmology

This result is too beautiful to be false; it is more important to have beauty in one's equations than to have them fit experiment.

P.A.M. Dirac on QED (Quantum Electrodynamics)[1]

In answer to the question of why it happened, I offer the modest proposal that our universe is simply one of those things which happen from time to time.

Edward P. Tryon [2]

The more the universe seems comprehensible, the more it also seems pointless.

Steven Weinberg

What if the Hokey-Pokey really *is* what it's all about?

Jimmy Buffett (more or less)

[1] Another Dirac quote, totally unrelated but too amusing not to include: "In science one tries to tell people, in such a way as to be understood by everyone, something that no one ever knew before. But in poetry, it's the exact opposite."

[2] Though expressed whimsically, this quote is actually serious; Tryon was the first to earnestly propose that the universe started as a quantum fluctuation. But since the mechanism of inflation had not yet been discovered, he had no explanation for how such a subatomic occurrence could become the size of the observed universe and so was not taken seriously at the time.

There are only two possibilities:

- The universe is a logical place.
- The universe is like an endless series of episodes of *South Park*.

Historically, physics has for the most part been based on the former premise.

In this chapter we will try to sketch some of the more significant developments in physics and to give you a very rough idea of their logical and historical place. While we will delve into a few topics here and there, for the most part this chapter is intended only as a skeletal outline; you can find a much fuller discussion in, among many others, the following books:[3]

Brian Greene, *The Elegant Universe*

> Greene gives a generally excellent discussion of all of physics, including a sort of pseudohistory, and emphasizing string theory and the corresponding cosmology. He arguably goes a bit over the top with the epic melodrama of his description of the discord between quantum theory and general relativity, and he gives short shrift to the hugely important topic of quantum field theory, but the book is otherwise a magnificent work. If you are only going to read one book about modern physics, this should be it.

Brian Greene, *The Fabric of the Cosmos*

> In this very thorough and excellent exposition, Greene traces the historical development of our understanding of space and time from Newton through string and M theory, including discussions of the arrow of time and of cosmology. While, because of its emphasis on the theme of space and time, the more angular perspective of *The Fabric of the Cosmos* does not provide as general an introduction to our current understanding of the physical universe as *The Elegant Universe*, this book is arguably the second book you should read if you are only going to read two books.

Brian Greene, *The Hidden Reality*

> In yet another thorough and excellent exposition, Greene discusses the amazing consequences of the various cosmological possibilities, from the classic infinite universe to the multiverse scenarios of eternal inflation, string theory, and other modern theories. While necessarily abbreviated in its presentation of the background physics, the text is nonetheless very lucid and has

[3] It is customary to cite the publisher, year of publication, etc., but we aren't here to do homage to a system of corporate greed.

a tremendous "wow factor." If you are only going to read three books about modern physics, this should be the third.

Leon Lederman, *The God Particle*

An ardent, even chauvinistic, experimentalist, Lederman openly biases his presentation toward experimental history and results, but he nonetheless does an excellent job of explaining collider/accelerator physics. His outline of the physics of the ancient Greeks is very good, and the whole book is written with an entertaining sense of humor. The title refers to the search for the Higgs particle.

Alan Guth, *The Inflationary Universe*

Guth gives a thorough and very lucid presentation of modern cosmology, with emphasis on inflation, a theory that he developed and arguably remains the foremost expert on. The book also has some interesting personal history that gives you an idea what it is like to be a theoretical physicist.

Lisa Randall, *Warped Passages: Unraveling the Mysteries of the Universe's Hidden Dimensions*

Because it restricts itself largely to those background aspects of relativity and quantum theory directly relevant to its titular thesis, *Warped Passages* lacks some of the "wow factor" of *The Elegant Universe*, but it makes an excellent followup to Greene for those interested in a more thorough discussion of technical aspects of the physics and of extra dimensions in particular. You can skip over the silly story with which each chapter begins; the rest is a remarkably clear and worthwhile exposition.

Steven Weinberg, *The First Three Minutes*

Weinberg's book is still quite valid in spite of its vintage (1977), and recent editions have an afterword with some discussion of what has been learned since it was first published. While his presentation might be a bit more technical than the average pedestrian would find palatable, Weinberg gives an excellent chronology of standard big-bang cosmology.

Stephen Hawking, *A Brief History of Time*, *The Universe in a Nutshell*, and *The Theory of Everything*

We recommend that you give these a miss. In our opinion, Hawking's books are marred by a tendency to ignore or gloss over

the work of others and to present Hawking's own work as more central and important than some would consider it.

A Brief History of Time does discuss some topics, such as the anthropic principle, not found in the other books listed here, and it also, in conjunction with parallel passages in Guth's *The Inflationary Universe*, affords an interesting study in prejudice and egotism in science, but it is rather dated now and is nowhere near as comprehensive as Greene's books.

The Universe in a Nutshell is also very good in some respects, but, with a presentation like that of a comic book, it is too cursory in its treatment of the various topics to constitute much more than a coffee-table, cocktail-party conversation piece.

Apparently just a typeset collection of lecture notes, *The Theory of Everything* not only is very skimpy, but does not, in spite of its title, even begin to talk about string theory until the last of its seven chapters.

22.1 The Early Days

The physics of the ancient Greeks was pretty much like modern economics or philosophy — you just made it up as you went along. It was all great fun, and everyone had a really good time, but there wasn't too much in the way of the scientific method — hypothesizing a theory, fleshing out the theory logically and mathematically and checking that it was sensible, working out the phenomenology (that is, the observable consequences of the theory), and then testing the theory's predictions experimentally.

We can also skip right over the Dark Ages: the Medievals devoted most of their time and energy to gadgets — mechanical devices to automate milling and grinding and so forth. They made no progress in our understanding of the physical universe and, one can't help but think, really couldn't have cared less about it. But at least they didn't spend their time worrying about college admissions.

The first of the major figures in physics to fully apply the modern scientific method was Galileo Galilei (1564-1642) with, among other ventures, his investigations of the kinematics of bodies rolling down inclines and of pendula.[4] Galileo carefully formulated and tested hypotheses by induction and deduction, drawing inferences based on observation and experimentation, then applying rigorous mathematical analysis. Galileo also made many astronomical observations that supported a heliocentric rather then a geo-

[4]Though there is still some debate about it, it seems pretty well established that Galileo did not make the fabled experiment of dropping balls from the Tower of Pisa.

centric model of the solar system.[5] As is often the case with people who have new ideas or different views, he was very nearly burned at the stake: people never allow facts to stand in the way of what they want to believe,[6] and Galileo was duly persecuted by those then in political power for challenging their irrational but comfortably self-serving beliefs.

22.2 Newton

Among the many accomplishments of Isaac Newton (1642-1727) were his mathematical formulation of dynamics and his inverse-square law of gravity, embodied in the formulæ

$$F = ma \qquad\qquad F = \frac{Gm_1m_2}{r^2}$$

with which you are by now so painfully familiar.

Although $F = ma$ accounted nicely for the motion of bodies and Newton's law of gravity predicted all three of Kepler's laws, it was simply assumed that the inertial mass in $F = ma$ was the same as the mass on which the gravitational force acted and that the gravitational force between the two masses m_1 and m_2 was communicated instantaneously across the distance r separating them (known as action at a distance) — assumptions for which there was no logical basis or ostensible mechanism.[7]

22.3 Maxwell & Others

In the course of the late eighteenth and the early nineteenth centuries, a great many people contributed to investigations of what seemed a great diversity of electric and magnetic phenomena. Gradually these at first apparently disparate phenomena were linked to one another. The great contribution of James Clerk Maxwell (1831-1879) was to see that *all* of electromagnetism is governed by just four equations.[8]

Maxwell's formulation of electromagnetism had several important features:

[5] Having trashed the whole dropping-balls-from-the-Tower-of-Pisa thing, we might as well note that, it seems, Galileo also did not invent the telescope, either. But he does get credit for making very good use of it.

[6] Just ask anyone in Kansas.

[7] Newton himself, as he noted in his *Principia*, was aware of these limitations in his work.

[8] Actually, the way Maxwell himself expressed the Maxwell equations was very messy; the equations were later re-expressed in their current simplified form by Oliver Heaviside.

- Electromagnetic forces were not instantaneously communicated. Rather, electric charge gave rise to an electric field that permeated all of space. Changes in this electric field (due, for example, to the motion of the charge) traveled outward from the charge at the finite speed of light, and the electric force felt by another charge depended only on the value of the electric field *at the location of that other charge* at the instant in question. The effect was entirely local; there was no "action at a distance." Similarly electric currents gave rise to magnetic fields, and it was these magnetic fields that exerted magnetic forces on other electric currents.

- Mathematically, the Maxwell equations have a wave solution that predicts not only the existence of light waves but also their speed of propagation: in terms of the fundamental electromagnetic constants ϵ_0 and μ_0 (the electric permittivity and magnetic permeability of the vacuum, respectively), the Maxwell equations give $c = 1/\sqrt{\epsilon_0\mu_0}$ for the speed of light. And since the Maxwell equations were derived without reference to any particular frame, one is led to conclude that the speed of light is independent of reference frame — a realization that later served as the basis of relativity.[9]

[9]Decades elapsed between the formulation of the Maxwell equations and special relativity, however, in good part because people had such a hard time swallowing the notion that the speed of light was independent of the reference frame. At first, people thought that there must be some hitherto unobserved medium, which they termed the æther, that carried light waves in the same way that the air carries sound waves. The result $c = 1/\sqrt{\epsilon_0\mu_0}$ would then hold only in the frame at rest relative to the æther; in all frames moving relative to the æther, the speed of light would be affected in the same way that the speed of sound is affected by the wind. In 1887, Albert Abraham Michelson and Edward Morley set up an experiment to detect the effect of such motion. Since the speed of light $c = 3.0 \times 10^8$ m/s is so large, motion of very high velocity was required; Michelson and Morley used the orbital velocity of the Earth around the Sun. While this velocity is still much smaller than the speed of light, it is large enough that any changes in the speed of light should have been observable. Mirrors and half-mirrors were used to split a light beam and bounce it back and forth along two paths, one parallel to the Earth's orbital motion and the other perpendicular to it. The effect of the Earth's motion through the æther on the light waves following these two paths was expected to be just like the headwind-tailwind and crosswind parts of problem #83 of Chapter 3 and should, when the waves were subsequently recombined, have resulted in a measurable shift in the interference pattern they produced. In fact, no effect beyond the bounds of experimental error was detected, neither in the original Michelson-Morley experiment nor in the many more accurate such experiments that have since been carried out. As much as it violated common sense, the only possible conclusion seemed to be that the speed of light is the same in all reference frames, regardless of their relative motion.

Those interested can find Michelson's and Morley's original paper at `http://www.aip.org/history/gap/PDF/michelson.pdf`.

- Maxwell's equations could be considered the first example of unification: seemingly widely disparate electromagnetic phenomena turn out to be just special cases of a relatively simple underlying principle. Moreover, the Maxwell equations are highly symmetric and could also be considered as the first hint of the symmetry principles that underlie all of modern physics.

22.4 Relativity

Working from the postulates that the speed of light and the laws of physics were the same in all inertial frames, Albert Einstein (1879-1955) reasoned out special relativity (1905): the invariant interval, time dilation, length contraction, $E = \gamma mc^2$, etc. That the speed of light is the same in all reference frames means that weird things happen to spatial distances and time intervals between frames. Many of these relativistic effects violate common sense; relativistic effects are small until one reaches speeds comparable to the speed of light, and our common sense is based on speeds far too low for them to be perceptible.

Not satisfied with the seemingly artificial restriction of the first postulate to just inertial frames, Einstein spent eleven years working out the mathematical consequences of relaxing this restriction and postulating simply that the laws of physics were the same in all frames, inertial or not. The result was general relativity — which, as it turned out, just happened to provide a theory of gravity. General relativistic gravitation reproduces familiar Newtonian gravity in the limit of low velocities and small masses — so that, like Newton's gravity, it predicts all of Kepler's laws and fully accounts for all commonly observed gravitational phenomena — , but it does not suffer from the pathologies of Newtonian gravity: in general relativity, gravitational effects travel only at the finite speed of light, so that there is no "action at a distance," and, since an acceleration due to a gravitational field should be physically identical to being in a frame that is accelerated nongravitationally, gravitational and inertial mass must be equal.[10] This equivalence of gravitational and inertial effects is known as the principle of equivalence.

[10]The classic example to illustrate this: experiencing 1 g of acceleration due to gravity in a nonaccelerating elevator on the surface of the Earth is physically identical to being in an elevator accelerating at 1 g far out in empty space (where no gravitational pull is exerted on it): you can slide objects down inclines, juggle balls, etc., and in no way be able to tell whether you are in the elevator on Earth or the elevator in space. (More precisely, gravitational and nongravitational accelerations are equivalent *locally*, meaning that at any point in spacetime there are equivalent frames. In our present example you could in fact distinguish the frames by comparing neighboring points: in the elevator in space, the accelerations of objects would all be in the same direction, while in the elevator on Earth the accelerations would all be toward the same point — the center of Earth — and therefore in slightly different directions.)

22.5 Quantum Mechanics

Atoms were known to consist of negatively charged electrons orbiting around positively charged nuclei, but their stability could not be explained: according to classical electromagnetism (that is, the Maxwell equations), the centripetal acceleration of the orbiting electrons should cause them to give off electromagnetic radiation in the form of light, with the result that they would lose energy and quickly spiral into the nucleus. Also unexplainable was the discreteness of atomic spectral lines: while it seemed that orbits of any energy should be possible, atoms were in fact found to have only certain discrete energy levels.

This difficulty was resolved by quantum mechanics, to which a great many people contributed, and the advent of which is often taken to coincide with the formulation of the wave equation for the hydrogen atom by Erwin Schrödinger (1887-1961) in 1926.

The basis of quantum mechanics is the wave function or probability amplitude. Previously there were thought to be continuous, wave-like forms of matter (such as light) and, distinct from these, discrete, particle-like forms (such as electrons). The dynamics of the former was governed by wave equations like those derived from the Maxwell equations, while that of the latter was governed by $F = ma$ or its relativistic equivalent. In either case, the dynamics was deterministic: one could, knowing the initial state of a wave or particle (for example, the initial location and velocity in the case of a particle), in principle calculate its motion exactly at all future and past times. But the quantum mechanical picture of physical reality was entirely different: all forms of matter were represented by probability amplitudes, from which one could calculate the *probability* that the matter was at a particular place at a particular time. Only these probability amplitudes could be determined exactly; exact locations or velocities could not be determined even in principle. Whether matter behaved like a continuous wave or a discrete particle depended on the circumstances: large quantities of a particular type of matter would behave like a continuous wave, small quantities like discrete particles. When, for example, a photograph is taken with very dim light, the "image" consists at first of just isolated dots, corresponding to the absorption of discrete particles of light (photons) by the film; with longer exposures, the dots eventually fill out a continuous image. Similarly, a beam consisting of large numbers of electrons will behave like a continuous wave — a property that makes it possible to use electrons in an electron microscope in the same way that light waves are used in an optical microscope.

When the wave solutions to quantum mechanical wave equations are solved for bound states such as atoms, it turns out that only certain dis-

crete solutions are allowed.[11] This explains the stability of atoms and the discreteness of atomic energy levels and spectra. Quantum mechanics was also able to account for a great number of other previously unexplainable phenomena, and one could include relativity in quantum mechanics as well (a formulation known, oddly enough, as relativistic quantum mechanics).

The statistical nature of physics posited by quantum theory bears great emphasis. In the early days of quantum mechanics, there were two warring factions: in one camp, Einstein, Schrödinger, Louis de Broglie, and others refused to accept the statistical nature of quantum physics, maintaining that in this respect quantum mechanics was just a "provisional" theory and that eventually we would discover a better theory that would restore determinism;[12] in the other camp (the "Copenhagen" camp), Werner Heisenberg, Niels Bohr, and others maintained that not only did the statistical nature of quantum mechanics accurately reflect the reality of nature, but that almost all classical notions should be abandoned. In 1935, in an effort to do quantum mechanics in, Einstein, Boris Podolsky, and Nathan Rosen proposed an experiment now known as the EPR paradox, in which a stationary particle would decay into two other particles that fly off (by conservation of momentum) in opposite directions. In the case that the decaying particle had zero spin angular momentum, the spin angular momenta of the two particles produced by the decay would (by conservation of angular momentum) have to add up to zero and thus be oriented in opposite directions. So if we call the two decay particles A and B, if A had spin "up," then B would have to be spin "down" and vice versa. The arguments made by Einstein, Podolsky, and Rosen are too involved and subtle to present here,[13] but the gist is that according to the probabilistic, quantum mechanical scheme the orientation of the spins of the decay particles were matters of probability; you couldn't be sure before making a measurement whether A had spin up or spin down. But as soon as you had measured the spin of A, you knew immediately the direction of the spin of B: it had to be opposite to that of A. To the Einstein trio, this nonlocal effect was a violation of causality: no information or physical effect can travel faster than the speed of light, and yet by making a measurement on A, you would instantaneously determine the state of B. Einstein, Podolsky, and Rosen rejected this as a "spooky" kind of action at a distance.

While parts of their reasoning could be argued to lie outside the domain of physics proper and to trespass into the domain of philosophy, the EPR paradox has since been resolved both theoretically and experimentally. In

[11] This will be shown explicitly in §22.5.2, in which the energy states of the infinite square well are worked out.

[12] You have probably heard Einstein's famous pronouncement, "Gott würfelt nicht" (God does not play dice).

[13] Einstein could be a weaselly wascal at times.

the view of Einstein, Podolsky, and Rosen, particles A and B existed in some definite state; our uncertainty about the orientation of their spins was a consequence, not of the statistical nature of the physical universe, but of the incompleteness of our understanding of physics and of our knowledge of the details of the decay process. In their scheme, there are other physical parameters, of which we are as yet unaware, that would restore determinism. Such schemes are known as hidden-variable theories. In 1964, John Bell was able to prove that the predictions made by quantum theory for the results of certain spin measurements in the EPR experiment differed in subtle but quantitatively measurable ways from those made by hidden-variable theories. Experimental measurement of these differences, which are known as Bell's inequalities, is, however, very difficult; it was not until 1982 that Alain Aspect succeeded in doing so. The experimental results fall squarely on the side of quantum theory; hidden-variable theories are ruled out.

This still leaves the issue of apparent nonlocality, that is, of a measurement on A resulting instantaneously in knowledge of the spin state of B. In the original Copenhagen interpretation of quantum mechanics, it was simply accepted that in this respect quantum mechanics was indeed nonlocal: before making the measurement of A's spin, the AB system was in a superposition of states, each with its own probability of occurrence, and the effect of the measurement on A is to filter this superposition of states and collapse the system's wave function down to a single, definite state, with this collapse occurring everywhere at the same time. For example, initially the AB system might have been a superposition of 25% A up and B down and 75% A down and B up. Repeated measurements on such a system would find, modulo statistical fluctuations, that 25% of the time A was up and B down, and the other 75% of the time A was down and B up. Any one measurement could go either way, but would definitely collapse down to one or the other definite state: A up and B down, or A down and B up. This whole business of wavefunction collapse has been rightfully regarded as extremely malodorous: no explanation is even suggested for why wave functions should collapse upon the making of a measurement — why shouldn't we, for example, be able to discover by a measurement that the system is in fact in a superposition of states? Moreover, this scheme makes a distinction between the observer (the person making the measurement) and the observed (the object of the measurement) without ever clearly defining what differentiates the observer from the observed. It would be more correct to regard the act of observation as an interaction, on equal footing, between the wave functions of two parties, each of whom is both observer and observed.

A wilder but logically sounder resolution of the nonlocality issue is provided by the many-worlds interpretation of quantum mechanics, first formulated in 1956 by Hugh Everett. While the details are again too involved and subtle to get into here, the gist of the many-worlds interpretation is that

there is no collapse of the wave functions; the system AB really does exist in a superposition of states, and when we make a measurement on the AB system our wave function becomes entangled with that of the AB system. In this scheme, all of the quantum possibilities are realized, each in its own separate universe. In the 25-75 example in the preceding paragraph, as a result of the measurement of the spin of A, the universe would split into two totally separate, noncommunicating universes, one in which we found that A was up and B down, and the other in which we found that A was down and B up. The many-worlds interpretation does not make any distinction between the observer and the observed, and it avoids the whole business of wave-function collapse, but the suggestion that the universe is continually branching out into zillions of independent descendant universes is, besides being outlandish, arguably trespassing into the domain of metaphysics: since those descendant universes cannot in any way communicate with or know about each other, there is no way to verify their existence. On the other hand, it could be argued that if a theory, after due application of Occam's razor, is otherwise in complete and compelling agreement with experiment, then any logically necessary but in principle untestable consequence of the theory can be regarded as equally well verified.

In the 1980s, another understanding of the measurement problem was proposed: decoherence. While we are again not in a position to get into the technical details, the basic premise of decoherence is that the system being observed, far from being isolated, continually undergoes myriad microscopic interactions with the environment — for example, exchanges of photons in the form of infrared thermal radiation — , and that consequently the calculation of quantum states and probabilities should properly be for the complete system consisting of the observer, the observed, *and* the environment. It turns out that the overall effect of the environment in such calculations does not simply factor out; rather, what would without the environmental interaction have been a coherent superposition of states is reduced to a set of incoherent (in the sense of what one might call discrete) outcomes, each with its own probability. This explains why, in the 25-75 example above, repeated measurements on the AB system find that 25% of the time the system is in the state A up/B down and 75% of the time in the state A down/B up, but no single measurement finds the system in a state that is a simultaneous superposition of 25% A up/B down and 75% A down/B up. While decoherence does explain why measurements don't discover systems to be in superpositions of states, it does not, however, claim to offer a broader understanding of quantum superposition, and in particular it does not address the issue of superposition for those those systems that are not immersed in environments — such as the universe itself.

To this day we have yet to arrive at a comprehensive, fundamental understanding of quantum states and probabilities. The great physicist Richard

Feynman famously asserted that no one really understands quantum mechanics. But Feynman is also famous for proposing what he called the "shut-up-and-calculate" interpretation of quantum mechanics:[14] once one becomes familiar with quantum theory, it is always clear enough how to calculate the states and probabilities, so that our lack of an understanding of their fundamental nature has, oddly enough, not prevented us from applying quantum theory with tremendous success: although technical problems have so far prevented us from quantizing gravity, we presently have quantum theories of the other three forces in nature (the electromagnetic, the strong, and the weak) that account extremely well for the physics we observe in the universe and are in agreement with experiment out to 11 decimal places. And whatever questions of interpretation of quantum theory remain, the issue of the statistical nature of the world in which we live has been resolved: it is not that the world merely seems probabilistic because our knowledge of it is incomplete, it is that it *is* probabilistic.

Some other significant features of quantum mechanics:

- In classical (that is, nonquantum) mechanics, an object cannot make it over a potential energy "bump" unless it has enough kinetic energy. For a cart on a roller coaster, for example, $E = K + U = \frac{1}{2}mv^2 + mgh$ means that the lowest value of v corresponds to the highest value of h, and since the lowest value the speed v can have is zero, there is an upper limit on how high the cart can go: if a bump in the track goes higher than this, the cart will come to rest before reaching the top of the bump and then roll back down; it cannot make it to the far side of the bump.

  In quantum mechanics, however, it is possible for the object to tunnel through the bump even when it doesn't have enough kinetic energy to make it over the top. It turns out that the wave-like probability amplitude, while it becomes extremely small, does not vanish in the bump, so there remains some small but nonzero chance that the object will end up on the far side of the bump.[15]

- In classical mechanics, there is no reason why all the various physical parameters of an object, like its mass, velocity, acceleration, location, momentum, etc., cannot be determined exactly and simultaneously. There will of course be limits to the accuracy with which these quantities can be measured in practice, but in principle every parameter can

[14] We should perhaps note that while the idea and the phrase *shut-up-and-calculate interpretation of quantum mechanics* is widely attributed to Feynman, apparently there is actually some uncertainty about its origins.

[15] Quantum tunneling will be dealt with quantitatively in §22.5.3, where a simple example will be worked out.

be determined exactly. Because of the wave-like nature of the probability amplitude and certain elementary properties of Fourier transforms (as the math involved is known), it turns out, however, that in quantum mechanics certain parameters are paired with each other and that the more exactly one is measured, the less exactly the other is known. One such pair is location and momentum (or, equivalently, velocity): the more precisely you know where the object is located, the less you know about how fast it is moving, and vice versa. In the extreme case that you know exactly where the object is, you will know nothing at all about how fast it is moving; and if you know its velocity exactly, you won't know anything about where it is located — it could be anywhere with equal probability.

The quantum mechanical principles governing these competing precisions or uncertainties are known as the Heisenberg uncertainty principles.[16]

Unfortunately, quantum mechanics was a dead end.[17] As a theory, it was always sprawling, messy, and incoherent, and it had no way of accounting for the great diversity of subatomic particles subsequently observed in collider experiments or the interactions between them. It was almost immediately superseded by quantum field theory, which will be the subject of §22.6. Before we get into field theory, however, we will delve into the details of some general features of quantum theories — wave functions and equations, probability amplitudes, the discreteness of bound states and energy levels, and quantum tunneling — , so that you can see for yourself how these effects come about. While our treatment will necessarily be rather elementary — if you go on in physics, your studies of quantum theory will extend over many years — , what we do in the pages ahead is the Real Thing.

22.5.1 Wave Functions & Operators

To get some more definite idea of what quantum mechanics and probability waves are all about, you first need to have a handle on traveling waves. Consider the behavior of

$$y = A\cos(kx - \omega t)$$

where A, k, and ω are constants, t is the time, and x and $y = y(x,t)$ are spatial coordinates. y might, for example, represent the vertical level of the

[16] Philosophers make a huge big deal out of the uncertainty principles and their implications for epistemology, but they frequently do not have even the most rudimentary understanding of them, with the result that they spout all sorts of amusing gibberish. Not that the result wouldn't have been the same in any event — philosophy could be defined as what happens when you take a good idea and get silly with it.

[17] For some odd reason, chemists have yet to notice this.

water in a water wave, with x corresponding to the horizontal location along the wave's direction of propagation. $A\cos(kx - \omega t)$ does in fact oscillate sinusoidally like a water wave, and those oscillations are both spatial and temporal: If we were to take a snapshot of the wave at some instant (that is, if we look at $y = A\cos(kx - \omega t)$ for some fixed value of t), the wave height varies essentially as $\cos kx$, so that, as expected, from a side view the profile of the wave is sinusoidal. If instead we fix our attention on the wave motion as a function of time at some particular location (that is, if we look at $y = A\cos(kx - \omega t)$ for some fixed value of x), the wave height varies essentially as $\cos \omega t$, so that, again as expected, the wave oscillates vertically up and down with the same sinusoidal motion as a harmonic oscillator.

As our notation implies, the constant ω is an angular velocity: since ωt is an angle, ω is an angular rate (angle/time). Now, as the time t changes by one period T, the wave should undergo one full cycle — going, say, from its highest point to its lowest and back again — , and in one full cycle the angle ωt should change by 2π. We therefore have $\omega T = 2\pi$ and hence

$$\omega = \frac{2\pi}{T} \tag{22.1}$$

just as we had for circular motion. Likewise we also have, in terms of the familiar frequency f,

$$f = \frac{1}{T} \qquad\qquad \omega = 2\pi f$$

The k in $y = A\cos(kx - \omega t)$ is known as the wave number and is the spatial analog of ω: a forward motion of one wavelength λ along the x direction (that is, a change of λ in the value of x) also corresponds to one full cycle, so that we have $k\lambda = 2\pi$ and hence

$$k = \frac{2\pi}{\lambda} \tag{22.2}$$

Just as the angular velocity ω gives a measure of how rapidly the wave is oscillating as a function of time, k gives a measure of how rapidly the wave is oscillating *spatially*: a wave with a higher value of ω has a shorter period and thus experiences more crests and troughs in a given interval of time; a wave with a higher value of k has a shorter wavelength, and thus more of its crests and troughs fit within a given distance.

Now suppose that we allow both x and t to vary, but fix our attention on one particular "point" on the wave, picking out, for example, a particular crest and watching it as it moves forward. Being at a fixed point in the cycle of a trigonometric function means being at a fixed value of the angle, so being at a crest or other fixed point on a wave means being at a fixed value of the angle $kx - \omega t$. In other words, as we watch the wave crest's forward progress there is no change in the angle $kx - \omega t$:

$$\Delta(kx - \omega t) = 0$$

Since only x and t change, this means that

$$k\Delta x - \omega\Delta t = 0$$

which in turn yields

$$\frac{\Delta x}{\Delta t} = \frac{\omega}{k}$$

Now, by definition $\Delta x/\Delta t$, the horizontal displacement of the wave divided by the corresponding time interval, is just the horizontal velocity of the point we are watching on the wave. Since ω and k are constant, so is the velocity $v = \omega/k$. Moreover, since the point we were watching could have been any point on the wave, this velocity applies to all points on the wave. We have thus established that $y = A\cos(kx - \omega t)$ represents a wave that not only oscillates sinusoidally in time and space, but that also moves forward (that is, propagates along the x axis) at constant velocity

$$v = \frac{\omega}{k} \tag{22.3}$$

It is as though the whole sinusoidal cross-sectional shape of the wave glides forward at this constant velocity. Using eqq. (22.1) and (22.2), eq. (22.3) can also be written

$$v = \frac{2\pi/T}{2\pi/\lambda} = \frac{\lambda}{T} = \lambda f \tag{22.4}$$

It turns out that in quantum mechanics, as in many other applications involving sinusoidal oscillations, expressions in terms of sines and cosines are cumbersome to work with. The problem is that differentiating a sine yields a cosine and vice versa. So instead one usually works with a complex exponential, which is essentially just a sine and cosine rolled together:

$$e^{i\theta} = \cos\theta + i\sin\theta$$

This sort of complex exponential has the nice property that (apart from any overall factors brought out by the chain rule) its derivative is itself. In quantum mechanics, a one-dimensional wave function ψ representing an oscillation like that of a water wave would therefore be expressed as

$$\psi = Ae^{i(kx-\omega t)}$$

Since wave functions like this are supposed to represent everything, including things we usually think of as particles, our next task is to figure out to what the ω and k correspond in the case of a wave function representing a particle.

The molecules in objects are in continual random thermal motion, and this random thermal jostling leads to the emission of electromagnetic radiation. This radiation can be absorbed by other molecules, which then re-emit

it, etc., etc. At a given temperature, the equilibrium to which this continual jostling, emission, and absorption settles down results in a steady spectrum (that is, frequency distribution) of thermal electromagnetic radiation that is known as the blackbody radiation spectrum. The hotter the object, the more electromagnetic radiation it gives off. At ordinary temperatures this radiation is emitted at frequencies in the infrared, well below those of the visible spectrum, but at higher temperatures it becomes plainly visible and is the reason that objects glow red- or white-hot. In 1899 Max Planck, in order to resolve certain difficulties in explaining the blackbody radiation spectrum, proposed that the energy of electromagnetic radiation, instead of being continuous, came in discrete amounts called quanta,[18] each quantum having energy hf, where f was the frequency of the radiation and h was a constant that today we know, oddly enough, as Planck's constant. This hypothesis led to a very good fit to the observed blackbody radiation spectrum with

$$h = 6.6260693 \times 10^{-34}\,\text{Js}$$

This discrete nature of the electromagnetic radiation that constitutes light was even more apparent in a later phenomenon known as the photoelectric effect,[19] which involves blowing electrons out of a piece metal by shining a light beam on it. This might not seem likely to a prove a very engaging or profitable pastime, but it turns out to reveal some very fundamental physics: While it takes a certain minimal amount of energy to blow each electron out of the metal, the energy density of an electromagnetic wave is proportional to the square of the wave's amplitude, so one would expect that light of any frequency would blow electrons out of the metal if only the beam were intense enough. In fact, one actually finds that there is a maximal value of the wavelength λ (or, phrased in terms of frequency, a minimal value of f) beyond which the beam cannot, no matter how intense, liberate electrons from the metal. The explanation of this phenomenon is that what seems like a continuous light wave is actually an aggregation of individual photons, each of energy hf; the more intense beams have more photons in them, but individual electrons are liberated from the metal by individual photons, and once the energy hf of the individual photons is too low to liberate electrons from the metal, it doesn't matter how many photons are in the beam — no electrons will be liberated. The photoelectric effect in fact turns out to be a simple way to measure Planck's constant h.

[18] This usage is in fact the origin of the *quantum* in *quantum mechanics*, *quantum field theory*, etc.

[19] It was actually for his explanation of the photoelectric effect, not for his theories of special or general relativity, that Einstein was awarded the Nobel prize 1921. Since Einstein went on to take great exception to quantum mechanics because of its statistical nature, this proved very ironic. But 1905 was a good year for Einstein: he published three papers — on the photoelectric effect, special relativity, and Brownian motion — , each of which was by itself more than worthy of a Nobel prize.

The natural supposition is that the same relation $E = hf$ that holds for photons also holds for other kinds of particles. Since $\omega = 2\pi f$, in terms of ω we have

$$E = hf = h\frac{\omega}{2\pi} = \frac{h}{2\pi}\omega$$

The combination $h/2\pi$ occurs so frequently in quantum theory that there is a special symbol for it: $\hbar$ (read "h bar"),

$$\hbar = \frac{h}{2\pi} = 1.05457168 \times 10^{-34}\ \text{Js}$$

Thus

$$E = \hbar\omega \tag{22.5}$$

Now, according to relativity theory, space and time are not separate; they transform into each other as one goes from one reference frame to another. So if the combination $kx - \omega t$ in the wave function ψ is to be meaningful, a relation similar to the $E = \hbar\omega$ that holds for the ω in the ωt term must hold for the k in the kx term. And since, as we saw in the chapter on relativity, the energy E is the time component of the momentum p, we expect that the relation we are looking for is[20]

$$p = \hbar k \tag{22.6}$$

We are thus led to conclude that, at least in the simple case of a wave function analogous to a water wave,[21] the wave function ψ of a particle is of the form

$$\psi = Ae^{i(kx-\omega t)} = A\exp i\left(\frac{p}{\hbar}x - \frac{E}{\hbar}t\right) = Ae^{i(px-Et)/\hbar} \tag{22.7}$$

Physically, ψ is interpreted as a probability amplitude, with the probability itself given by $|\psi|^2$. Since ψ is complex, the absolute square of ψ is given by

$$|\psi|^2 = \psi\psi^*$$

where ψ^* is the complex conjugate of ψ. More precisely, $|\psi|^2$ is a probability density, meaning that the probability of finding the particle somewhere between x and $x + dx$ is

$$|\psi|^2\, dx$$

[20] So hypothesized de Broglie in 1923, though more commonly you see this relation written, using eq. (22.2), in the equivalent form $p = h/\lambda$:

$$p = \hbar k = \frac{h}{2\pi}\frac{2\pi}{\lambda} = \frac{h}{\lambda}$$

[21] In technical terms, we are talking about a plane wave, that is, a wave with a straight wave front, as opposed to the cylindrical waves produced by tossing a pebble into a pond or the spherical sound waves that emanate from a point source.

with the probability of being found in the finite region from, say, x_1 to x_2 given by

$$\int_{x_1}^{x_2} |\psi|^2 \, dx$$

For our present purposes, the key thing to note is that when working with the wave function ψ, the momentum p and energy E can be expressed in terms of derivatives of ψ: applying $\partial/\partial x$ and $\partial/\partial t$ to ψ yields

$$\frac{\partial \psi}{\partial x} = \frac{\partial}{\partial x} A e^{i(px-Et)/\hbar} = \frac{i}{\hbar} p A e^{i(px-Et)/\hbar} = \frac{i}{\hbar} p \psi$$
$$\frac{\partial \psi}{\partial t} = \frac{\partial}{\partial t} A e^{i(px-Et)/\hbar} = -\frac{i}{\hbar} E A e^{i(px-Et)/\hbar} = -\frac{i}{\hbar} E \psi$$

Turning these relations around, and remembering that $1/i = -i$, we have

$$p\psi = -i\hbar \frac{\partial \psi}{\partial x} \qquad\qquad E\psi = i\hbar \frac{\partial \psi}{\partial t}$$

so that, when acting on ψ, p and E are the differential operators

$$p = -i\hbar \frac{\partial}{\partial x} \qquad\qquad E = i\hbar \frac{\partial}{\partial t} \tag{22.8}$$

With these expressions for momentum and energy, we can derive and solve the Schrödinger equation in the next subsection.

22.5.2 The Schrödinger Equation & Discrete States

In Newtonian mechanics, the energy E of a particle was given by the sum of its kinetic and potential energies: $E = K + U$. To generate the quantum mechanical equivalent, we simply use $p = mv$ and $p = -i\hbar\,\partial/\partial x$ to rewrite

$$K = \tfrac{1}{2}mv^2 = \frac{(mv)^2}{2m} = \frac{p^2}{2m} = \frac{1}{2m}\left(-i\hbar \frac{\partial}{\partial x}\right)^2 = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}$$

Thus $E = K + U$ becomes the time-independent Schrödinger equation:[22]

$$E\psi = -\frac{\hbar^2}{2m}\frac{\partial^2 \psi}{\partial x^2} + U\psi \tag{22.9}$$

[22] The time-dependent Schrödinger equation is obtained by further noting that $E = i\hbar\,\partial/\partial t$, so that eq. (22.9) can also be written

$$i\hbar \frac{\partial \psi}{\partial t} = -\frac{\hbar^2}{2m}\frac{\partial^2 \psi}{\partial x^2} + U\psi$$

If we know the potential energy U corresponding to the force acting on a particle, we can use eq. (22.9) to determine both the particle's wave function ψ and its possible energies E.

To determine the states of the hydrogen atom, for example, we first need to extend eq. (22.9) to three dimensions by including also the y and z directions:

$$E\psi = -\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2}\right) + U\psi = -\frac{\hbar^2}{2m}\nabla^2\psi + U\psi \qquad (22.10)$$

where ∇^2 is the Laplace operator

$$\begin{aligned}\nabla^2 &= \nabla\cdot\nabla \\ &= \left(\frac{\partial}{\partial x}\,\hat{\mathbf{x}} + \frac{\partial}{\partial y}\,\hat{\mathbf{y}} + \frac{\partial}{\partial z}\,\hat{\mathbf{z}}\right)\cdot\left(\frac{\partial}{\partial x}\,\hat{\mathbf{x}} + \frac{\partial}{\partial y}\,\hat{\mathbf{y}} + \frac{\partial}{\partial z}\,\hat{\mathbf{z}}\right) \\ &= \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\end{aligned}$$

For the hydrogen atom, the potential energy U is just the electrostatic potential energy

$$U = \frac{1}{4\pi\epsilon_0}\frac{q_1 q_2}{r}$$

with q_1 being the electron charge $-e$ and q_2 being the proton charge $+e$:

$$U = -\frac{1}{4\pi\epsilon_0}\frac{e^2}{r}$$

Thus the time-independent Schrödinger equation for the hydrogen atom is

$$E\psi = -\frac{\hbar^2}{2m}\nabla^2\psi - \frac{1}{4\pi\epsilon_0}\frac{e^2}{r}\,\psi$$

The object of the game is to solve this equation, in spherical coordinates (r,θ,φ), for the wave function $\psi(r,\theta,\varphi)$. Unfortunately, the math required to accomplish this is beyond what we can get into here, but the solution turns out to be

$$\psi(r,\theta,\varphi) = -\sqrt{\left(\frac{2}{na}\right)^3\frac{(n-\ell-1)!}{2n\Big((n+\ell)!\Big)^3}}\;e^{-r/na}\left(\frac{2r}{na}\right)^{\ell} L_{n+\ell}^{2\ell+1}\left(\frac{2r}{na}\right) Y_m^{\ell}(\theta,\varphi)$$

where a is the Bohr radius

$$a = 4\pi\epsilon_0\frac{\hbar^2}{me^2} = 0.5291772108\times 10^{-10}\,\mathrm{m}$$

L is an associated Laguerre polynomial defined by

$$L_{n+\ell}^{2\ell+1}(\xi) = \sum_{k=0}^{n-\ell-1} (-1)^{k+1} \frac{\Big((n+\ell)!\Big)^2}{(n-\ell-1-k)!(2\ell+1+k)!}\, \xi^k$$

Y_m^ℓ is the spherical harmonic

$$Y_m^\ell(\theta,\varphi) = \sqrt{\frac{2\ell+1}{4\pi}\frac{(\ell-m)!}{(\ell+m)!}}\, P_m^\ell(\cos\theta)\, e^{im\varphi}$$

P_m^ℓ is the associated Legendre polynomial

$$P_m^\ell(\xi) = \frac{(-1)^m}{2^\ell \ell!}\,(1-\xi^2)^{\frac{m}{2}}\,\frac{d^{\,\ell+m}}{d\xi^{\ell+m}}(\xi^2-1)^\ell$$

and n, ℓ, and m are integers such that

$$n = 1,\, 2,\, 3,\, 4,\, \ldots \qquad \ell = 0,\, 1,\, 2,\, \ldots,\, n-1 \qquad |m| \le \ell$$

For a given set of values of n, ℓ, and m, the corresponding value of the energy E turns out to be

$$E = -\tfrac{1}{2}mc^2\alpha^2\,\frac{1}{n^2} = -13.6056923\,\text{eV}\,\frac{1}{n^2}$$

where α is the fine-structure constant

$$\alpha = \frac{e^2}{4\pi\epsilon_0\hbar c} = 7.297352568\times10^{-3} = \frac{1}{137.03599911}$$

Note that only certain discrete values of the energy are possible; for other values, there is no solution to the time-independent Schrödinger equation. The values of n, ℓ, and m correspond to the orbitals you learned about in chemistry: $n = 0$ is the ground state (the state of lowest energy), $n = 1$ the first excited state, $n = 2$ the second excited state, and so on, with $\ell = 0,\, 1,\, 2,\, 3,\, \ldots$ corresponding to the s, p, d, f, $\ldots$ orbitals.

At least, these are the solutions for the case $E < 0$ corresponding to electron kinetic energies insufficient to escape the electric pull of the nucleus — that is, for states such that $|K| < |U|$, so that $K + U < 0$. These are known as the bound states. For $E \ge 0$, the electron has enough kinetic energy to completely escape the pull of the nucleus, and such states, known as continuum states, have a different solution for the wave function ψ and can, as it turns out, have any (positive) value of E.

Although we are not in a position to solve the case of the hydrogen atom, we can work through a simpler, one-dimensional case that will illustrate how

only certain discrete bound states end up being allowed. Consider the infinite square-well potential

$$U(x) = \begin{cases} 0 & 0 < x < \ell \\ \infty & x < 0 \text{ or } x > \ell \end{cases}$$

Such a potential energy corresponds to zero force on the interior $0 \leq x \leq \ell$ of the well and at its walls to an infinite force that confines the particle, and hence the wave function ψ, to the region $0 < x < \ell$. Since $U = 0$ for $0 < x < \ell$, the time-independent Schrödinger equation within that region is, from eq. (22.9),

$$E\psi = -\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2}$$

which we can rewrite as

$$\frac{d^2\psi}{dx^2} = -\frac{2mE}{\hbar^2}\psi \tag{22.11}$$

We want to solve this equation for $\psi(x)$ and, in the process, to see what restrictions there may be on the possible values of E. To do this, we note that eq. (22.11) is identical to that of a mass on the end of an ideal spring, which was

$$F = ma$$
$$-kx = m\frac{d^2x}{dt^2}$$

and hence

$$\frac{d^2x}{dt^2} = -\frac{k}{m}x \tag{22.12}$$

The solution for the mass on a spring, as worked out in Chapter 9, was

$$x = A\cos(\omega t + \varphi)$$

where A and φ were arbitrary constants and ω was given by

$$\omega = \sqrt{\frac{k}{m}}$$

With the correspondences

$$\left.\begin{matrix} x \\ t \\ k/m \end{matrix}\right\} \leftrightarrow \left\{\begin{matrix} \psi \\ x \\ 2mE/\hbar^2 \end{matrix}\right.$$

between the spring equation (22.12) and the wave-function equation (22.11), we see that the solution for ψ will be

$$\psi = A\cos\left(\frac{2mE}{\hbar^2}x + \varphi\right) \tag{22.13}$$

Now, since the particle must have zero probability of being found at $x \leq 0$ and $x \geq \ell$, we must have $\psi(0) = 0$ and $\psi(\ell) = 0$:

$$0 = \psi(0) = A\cos\varphi$$
$$0 = \psi(\ell) = A\cos\Big(\frac{2mE}{\hbar^2}\ell + \varphi\Big)$$

From the former of these conditions, we must, since $A = 0$ would be very silly, have $\varphi = \frac{\pi}{2}$ or some equivalent angle. We will choose $\varphi = \frac{3}{2}\pi$ simply for convenience, so that the latter of the above conditions becomes

$$0 = A\cos\Big(\frac{2mE}{\hbar^2}\ell + \tfrac{3}{2}\pi\Big) = A\sin\Big(\frac{2mE}{\hbar^2}\ell\Big)$$

The angle of which we are taking the sine must therefore be an integer multiple of π:[23]

$$\frac{2mE}{\hbar^2}\ell = n\pi \qquad\qquad n = 1,\, 2,\, 3,\, \ldots$$

The possible energies are of the particle are thus discrete:

$$E = \frac{\pi\hbar^2}{2m\ell}\,n \qquad\qquad n = 1,\, 2,\, 3,\, \ldots$$

If we label these allowed energies E_n, our solution (22.13) for ψ becomes

$$\psi = A\cos\Big(\frac{2mE_n}{\hbar^2}x + \tfrac{3}{2}\pi\Big) = A\sin\Big(\frac{2mE_n}{\hbar^2}x\Big)$$

What determines the value of A? Remember that

$$\int_{x_1}^{x_2} |\psi|^2\,dx$$

is the probability of finding the particle in the region $x_1 \leq x \leq x_2$. Since the particle must be somewhere between $x = 0$ and $x = \ell$ with probability 1 (that is, 100%), we have

$$1 = \int_0^\ell \Big|A\sin\Big(\frac{2mE_n}{\hbar^2}x\Big)\Big|^2\,dx = \int_0^\ell |A|^2\sin^2\Big(\frac{2mE_n}{\hbar^2}x\Big)\,dx$$

which, if we plug in the value of E_n, reduces to

$$1 = \int_0^\ell |A|^2\sin^2\left(\frac{2m}{\hbar^2}\frac{\pi\hbar^2 n}{2m\ell}x\right)dx = |A|^2\int_0^\ell \sin^2\Big(\frac{n\pi x}{\ell}\Big)\,dx$$

[23] We do not need to allow zero or negative values of n because m, E, ℓ, and $\hbar$ are all positive.

With the change of variables

$$u = \frac{x}{\ell} \qquad\qquad dx = \ell\, du$$

this further simplifies to

$$1 = |A|^2 \int_0^1 \sin^2(n\pi u)\ell\, du = \ell\, |A|^2 \int_0^1 \sin^2(n\pi u)\, du$$

You can do out this integral, or simply remember that the average value of $\sin^2$ or $\cos^2$ over any whole number of half-cycles is $\frac{1}{2}$:

$$1 = \ell\, |A|^2\, \tfrac{1}{2}$$

Thus we arrive at

$$|A| = \sqrt{\frac{2}{\ell}} \tag{22.14}$$

Remember now that A is a complex number, so that eq. (22.14) is telling us that A is somewhere on the perimeter of a circle of radius $\sqrt{2/\ell}$ in the complex plane. If we were combining the wave functions of two or more particles, the phase of A (that is, its position on the circle) would make a difference, but for a single particle it is irrelevant, in the sense that the phase has no effect on the value of $|\psi|^2$ or anything else physically observable. We may therefore simply choose

$$A = \sqrt{\frac{2}{\ell}}$$

The solution for ψ at which we finally arrive is thus

$$\psi = \sqrt{\frac{2}{\ell}} \sin\left(\frac{2mE_n}{\hbar^2}x\right)$$

The value of A is called a normalization factor; normalization factors simply guarantee that the probability of the particle being somewhere in space is 1. Within the region $0 \le x \le \ell$, the probability of finding the particle between, say, x_1 and x_2 is then

$$\int_{x_1}^{x_2} |\psi|^2\, dx = \int_{x_1}^{x_2} \frac{2}{\ell} \sin^2\left(\frac{2mE_n}{\hbar^2}x\right)\, dx$$

Not that this would be terribly illuminating to work out; we're just trying to be clear about the interpretation of our solution for the wave function ψ in terms of probability.

22.5.3 Quantum Tunneling

As we saw just above in §22.5.1, the dependence of the wave function ψ of a particle on its momentum p and energy E is

$$e^{i(px-Et)/\hbar}$$

Suppose now that you, mass m, are standing at rest beside a brick wall of height H and thickness ℓ. Your kinetic energy is $K = 0$, and, if we put $y = 0$ at the level of your center of mass, your gravitational potential energy $U = mgy_{cm} = 0$ and total energy $E = K + U = 0$ vanish as well. Not a very interesting situation, it would seem.

In Newtonian mechanics, this is pretty much the end of the story; you can't make it over the wall not only because you are not moving, but also because you do not have enough energy to make it over the gravitational potential energy bump mgH presented by the wall. But in quantum mechanics this is not a concern. Forging ahead undaunted,[24] we note that your energy E "inside" the wall would be

$$E = \tfrac{1}{2}mv^2 + mgy = \frac{p^2}{2m} + mgH$$

Since, as we just noted, your energy E vanishes, this means that inside the wall your momentum is imaginary:

$$0 = \frac{p^2}{2m} + mgH$$

yields

$$p = \sqrt{-2m^2gH} = im\sqrt{2gH}$$

The effect on the momentum part $e^{ipx/\hbar}$ of your wave function as it goes "through" the wall would therefore be given by the factor

$$e^{ipx/\hbar} = e^{i(im\sqrt{2gH})\ell/\hbar} = e^{-m\ell\sqrt{2gH}/\hbar}$$

That is, if the value of your wave function were essentially $\psi = 1$ on your side of the wall, corresponding to virtual certainty that that is where you are located, your wave function would be approximately

$$\psi \approx 1 \cdot e^{-m\ell\sqrt{2gH}/\hbar} = e^{-m\ell\sqrt{2gH}/\hbar}$$

on the far side of the wall. The probability of your being found on the far side of the wall would therefore be approximately

$$|\psi|^2 \approx \left|e^{-m\ell\sqrt{2gH}/\hbar}\right|^2 = e^{-2m\ell\sqrt{2gH}/\hbar}$$

[24] Some of you will be familiar with this procedure from exams, where you have forged ahead even in Newtonian calculations in spite of having obtained imaginary results for momentum, velocity, etc.

For a mass m of, say, 60 kg, with a wall of height $H = 2$ m and thickness $\ell = \frac{1}{4}$ m, this works out to very roughly

$$e^{-2\times 10^{36}}$$

This should strike you as rather small. In fact, you would have to wait zillions of times the total lifetime of the universe to have a decent chance of being spontaneously found on the far side of the wall. But there are three significant points to be made: First, however small, the probability is nonzero: in the quantum world — and we do live in a quantum world — it *is* possible that you could tunnel through the wall and be found on the far side. Second, the probability of tunneling is much higher for smaller masses and potential energy bumps, such as those that occur on subatomic scales. Tunneling would in fact have been the explanation for cold fusion (§12.4) had it actually occurred. The idea was that deuterium nuclei, which normally have to be heated up to enormously high thermal kinetic energies in order to overcome their mutual electrical repulsion and fuse into helium, could have tunneled through this electrical potential energy barrier at significant rates even at room temperature if enough of them were simply kept close enough together. Unfortunately, while in principle this is possible, in practice cold fusion has not yet been induced. Finally, third, you are not atomic in the sense of being indivisible — you are made up of myriad smaller particles, and it is far more likely (though still highly improbable) that a single electron from you would tunnel through the wall than that all of you would do so simultaneously. And even if by chance all the particles of which you are composed did tunnel through simultaneously, it is very unlikely that on the far side of the wall they would form a coherent aggregation anything like yourself. It would be a major bummer.

22.5.4 The Quantum Harmonic Oscillator

Because of its fundamental importance in physics and the big deal we've consequently made of it elsewhere in the text, in this section we will determine the energy spectrum of the quantized harmonic oscillator, that is, its possible energy states. Unfortunately, while the results we will obtain are very simple, there isn't really any way we can motivate the steps in the derivation without presenting an inordinate amount of background material, so you pretty much just have to enjoy the ride.

The big-people symbol for the total energy $K + U$ is H, where H stands for the Hamiltonian. Although the symbol and the term are new, H is just the familiar total energy we formerly denoted by E, only now you get to feel really cool by writing it as H and using jargon like *Hamiltonian*. Anyway,

using $p = mv$, we can write the kinetic energy in terms of the momentum p:

$$K = \tfrac{1}{2}mv^2 = \frac{(mv)^2}{2m} = \frac{p^2}{2m}$$

If now we recall that the angular frequency for a mass m on the end of a spring of spring constant k is $\omega = \sqrt{k/m}$, we have

$$H = K + U = \frac{p^2}{2m} + \tfrac{1}{2}kx^2 = \frac{p^2}{2m} + \tfrac{1}{2}m\omega^2x^2$$

Now, $\hbar\omega$ has, if you work it out, the dimensions of energy, and it is in fact the only quantity you can construct out of the parameters m and ω and fundamental constants like $\hbar$ and c that has energy dimensions. So if we pull out an overall factor of $\hbar\omega$ —

$$H = \hbar\omega\left(\frac{1}{2m\hbar\omega}(p^2 + m^2\omega^2x^2)\right) \tag{22.15}$$

— , then the quantity within the outer parentheses is dimensionless. This will prove convenient shortly.

Now comes the trick: although it is not something that would occur to you without either considerably more study of these sorts of things or the ingestion of wildly irresponsible combinations of psychedelic drugs, we can define a new animal a by

$$a = \frac{p - im\omega x}{\sqrt{2m\hbar\omega}} \tag{22.16}$$

The Hermitian conjugate of a is denoted by $a^\dagger$ and is, in matrix terms, the complex conjugate of the transpose of a. Here we're not dealing with a and $a^\dagger$ as matrices, so we have simply the complex conjugate:

$$a^\dagger = \frac{p + im\omega x}{\sqrt{2m\hbar\omega}} \tag{22.17}$$

where we have noted that m, $\hbar$, ω, p, and x are all real. The product $a^\dagger a$ is thus

$$\begin{aligned} a^\dagger a &= \left(\frac{p + im\omega x}{\sqrt{2m\hbar\omega}}\right)\left(\frac{p - im\omega x}{\sqrt{2m\hbar\omega}}\right) \\ &= \frac{1}{2m\hbar\omega}\left(p^2 + im\omega xp - im\omega px + m^2\omega^2x^2\right) \end{aligned} \tag{22.18}$$

You might think that the cross terms simply cancel each other out, but here you have to remember that since we are doing a quantum calculation, p is not a simple variable; it is, by eqq. (22.8), the differential operator

$$p = -i\hbar\frac{d}{dx}$$

So when p and x are applied to a wave function ψ, the derivative in p will act on everything to its right and we have, using the product rule,

$$\begin{aligned}(px - xp)\psi &= \left(\left(-i\hbar\frac{d}{dx}\right)x - x\left(-i\hbar\frac{d}{dx}\right)\right)\psi \\ &= -i\hbar\frac{d}{dx}(x\psi) + x\,i\hbar\frac{d\psi}{dx} \\ &= \left(-i\hbar\frac{dx}{dx}\psi - i\hbar\,x\,\frac{d\psi}{dx}\right) + x\,i\hbar\frac{d\psi}{dx} \\ &= -i\hbar\psi\end{aligned}$$

Since this relation applies for any wave function ψ, it is often more succinctly written as

$$[p, x] = -i\hbar$$

where $[p, x]$, called the commutator of p and x, is defined to be [25]

$$[p, x] = px - xp$$

So the cross terms in eq. (22.18) do not cancel; rather, we have

$$\begin{aligned}a^\dagger a &= \frac{1}{2m\hbar\omega}\left(p^2 - im\omega(px - xp) + m^2\omega^2x^2\right) \\ &= \frac{1}{2m\hbar\omega}\left(p^2 - im\omega[p, x] + m^2\omega^2x^2\right) \\ &= \frac{1}{2m\hbar\omega}\left(p^2 - im\omega(-i\hbar) + m^2\omega^2x^2\right) \\ &= \frac{1}{2m\hbar\omega}\left(p^2 - m\hbar\omega + m^2\omega^2x^2\right) \\ &= \frac{1}{2m\hbar\omega}\left(p^2 + m^2\omega^2x^2\right) - \tfrac{1}{2}\end{aligned} \tag{22.19}$$

Comparing this result with eq. (22.15), we see that the oscillator energy can be written

$$H = \hbar\omega\left(a^\dagger a + \tfrac{1}{2}\right) \tag{22.20}$$

At this point, the question uppermost in your mind is probably, So what? Well, we are trying to determine the possible energy states of the quantum

[25] More generally, $[a, b] = ab - ba$ for any quantities a and b, not just p and x. It turns out that only when the commutator of two physical quantities vanishes can the values of those two quantities be simultaneously specified or known. The nonvanishing of the commutator of p and x corresponds to the Heisenberg uncertainty principle and the impossibility of simultaneously determining the exact values of both a particle's momentum and its location (or, equivalently, its velocity and location).

harmonic oscillator, and we have now written the oscillator's energy in terms of a and $a^\dagger$. So if we can figure out the properties of a and $a^\dagger$, we will be able to determine the oscillator's spectrum. The first thing to note is that the value of $a^\dagger a$ when applied to an oscillator state must be nonnegative. To see this, let us denote the value of $a^\dagger a$ on the state ψ by q: $a^\dagger a\psi = q\psi$. Since the probability

$$\int dV\, |\psi|^2$$

of finding the mass m *somewhere* in all of space has to be 1,

$$\int dV\, \psi^* a^\dagger a\psi = \int dV\, \psi^* q\psi = q\int dV\, \psi^*\psi = q\int dV\, |\psi|^2 = q(1) = q$$

But we can also gerrymander this integral into the form

$$\int dV\, \psi^* a^\dagger a\psi = \int dV\, (a\psi)^\dagger a\psi = \int dV\, |a\psi|^2$$

Since this last integral is of an absolute square, it must be nonnegative. From this we draw several conclusions: First, the value q of $a^\dagger a$ must be nonnegative. Second, in order to have $q = 0$, we must have $a\psi = 0$, that is, the operator a must annihilate the wave function ψ that corresponds to $q = 0$. Third, the nonnegativity of $a^\dagger a$ means that our result (22.20) for the energy has a minimal possible value:

$$H = \hbar\omega\Big(a^\dagger a + \tfrac{1}{2}\Big) \geq \tfrac{1}{2}\hbar\omega$$

The lowest possible value, $H = \frac{1}{2}\hbar\omega$, corresponds to the state ψ for which $q = 0$, that is, for which $a\psi = 0$. We will call this lowest-energy state the ground state and denote it by ψ_0.

Next we look at some more commutators. First, the commutator of a and $a^\dagger$:

$$[a, a^\dagger] = aa^\dagger - a^\dagger a$$

Just above (in eq. (22.19)) we worked out that

$$a^\dagger a = \frac{1}{2m\hbar\omega}\Big(p^2 + m^2\omega^2 x^2\Big) - \tfrac{1}{2}$$

If we were to do an exactly similar calculation of $aa^\dagger$, we would find that

$$aa^\dagger = \frac{1}{2m\hbar\omega}\Big(p^2 + m^2\omega^2 x^2\Big) + \tfrac{1}{2}$$

Subtracting the result for $a^\dagger a$ from that for $aa^\dagger$, we arrive at

$$[a, a^\dagger] = aa^\dagger - a^\dagger a = 1 \tag{22.21}$$

For the commutator of H and a, we then have

$$\begin{aligned}[H,a] &= Ha - aH \\ &= \hbar\omega\Big(a^\dagger a + \tfrac{1}{2}\Big)\,a - a\hbar\omega\Big(a^\dagger a + \tfrac{1}{2}\Big) \\ &= \hbar\omega\Big(a^\dagger aa + \tfrac{1}{2}a - aa^\dagger a - \tfrac{1}{2}a\Big) \\ &= \hbar\omega(a^\dagger aa - aa^\dagger a)\end{aligned}$$

With a little gerrymandering, this can be written in terms of $[a, a^\dagger]$ and thus simplified:

$$[H,a] = -\hbar\omega(-a^\dagger a + aa^\dagger)a = -\hbar\omega[a,a^\dagger]a = -\hbar\omega(1)a = -\hbar\omega a \tag{22.22a}$$

By exactly similar gerrymandering, we also have

$$[H,a^\dagger] = Ha^\dagger - a^\dagger H = \hbar\omega a^\dagger \tag{22.22b}$$

All of this probably still has a very large "so what" factor, but eqq. (22.22a) and (22.22b) are in fact telling us that the operators $a^\dagger$ and a have the effect of raising and lowering the energy of a state, respectively. To see this, suppose we act with $Ha^\dagger - a^\dagger H = \hbar\omega a^\dagger$ on a state $\tilde\psi$ of energy $\tilde E$ (that is, a state such that $H\tilde\psi = \tilde E\tilde\psi$):

$$\begin{aligned}(Ha^\dagger - a^\dagger H)\tilde\psi &= \hbar\omega a^\dagger\tilde\psi \\ Ha^\dagger\tilde\psi - a^\dagger H\tilde\psi &= \hbar\omega a^\dagger\tilde\psi \\ Ha^\dagger\tilde\psi - a^\dagger\tilde E\tilde\psi &= \hbar\omega a^\dagger\tilde\psi \\ Ha^\dagger\tilde\psi &= (\tilde E + \hbar\omega)a^\dagger\tilde\psi\end{aligned}$$

In other words, the value of H for the state $a^\dagger\tilde\psi$ is $\tilde E + \hbar\omega$ — higher by $\hbar\omega$ than its value for the state $\tilde\psi$. Similarly the energy of the state $a\tilde\psi$ is lower than that of the state $\tilde\psi$ by $\hbar\omega$. For this reason, the operators $a^\dagger$ and a are called raising and lowering operators (or sometimes ladder operators).

This property of the operators $a^\dagger$ and a is the reason why a annihilates the ground state ψ_0: no state of lower energy exists. It also tells us the spectrum of the harmonic oscillator: we construct the states of higher energy by acting on the ground state ψ_0 with $a^\dagger$ one or more times: each time $a^\dagger$ acts, the energy of the state increases by $\hbar\omega$. The state ${a^\dagger}^n\psi_0$ will thus have energy $(n + \frac{1}{2})\hbar\omega$.

In fact, the operator $a^\dagger a$ that occurs in H is called the number operator because of its effect on the state ${a^\dagger}^n\psi_0$: if we make use of eq. (22.21), we have

$$\begin{aligned}a^\dagger a\Big({a^\dagger}^n\psi_0\Big) &= a^\dagger a{a^\dagger}^n\psi_0 \\ &= a^\dagger(aa^\dagger){a^\dagger}^{n-1}\psi_0\end{aligned}$$

$$
\begin{aligned}
&= a^\dagger(aa^\dagger - a^\dagger a + a^\dagger a){a^\dagger}^{n-1}\psi_0 \\
&= a^\dagger([a, a^\dagger] + a^\dagger a){a^\dagger}^{n-1}\psi_0 \\
&= a^\dagger(1 + a^\dagger a){a^\dagger}^{n-1}\psi_0 \\
&= {a^\dagger}^n\psi_0 + {a^\dagger}^2 a{a^\dagger}^{n-1}\psi_0
\end{aligned}
$$

In other words, each time we shift the a to the right of an $a^\dagger$, we get an extra contribution of ${a^\dagger}^n\psi_0$. Shifting a to the right of all n factors of $a^\dagger$ will therefore give us n such extra contributions:

$$a^\dagger a\Big({a^\dagger}^n\psi_0\Big) = n{a^\dagger}^n\psi_0 + {a^\dagger}^n a\psi_0$$

But since ψ_0 is annihilated by a, the second term vanishes and we have simply

$$a^\dagger a\Big({a^\dagger}^n\psi_0\Big) = n{a^\dagger}^n\psi_0$$

In other words, when acting on the state ${a^\dagger}^n\psi_0$ the value of the operator $a^\dagger a$ is n.

Pulling everything together, what we have is this: the energy of the quantum harmonic oscillator can be expressed as

$$H = \hbar\omega\Big(a^\dagger a + \tfrac{1}{2}\Big)$$

which turns out to take only the discrete values

$$E = \hbar\omega\Big(n + \tfrac{1}{2}\Big) \qquad\qquad n = 0,\ 1,\ 2,\ 3,\ \ldots$$

The ground state is $n = 0$, for which $E = \frac{1}{2}\hbar\omega$, and the states $n > 0$ are the excited states of the oscillator. In quantum mechanics, that's pretty much the end of the story. But in quantum field theory, it turns out that a quantized field is like an infinite set of harmonic oscillators, one at each point in spacetime. The values of the field at each point are, like the states of an oscillator, discrete, with n corresponding to the number of particles (of whatever type is associated with the field) at that point. The effect of the raising and lowering operators $a^\dagger$ and a in quantum field theory is thus to create or destroy particles, and these operators are therefore called creation and annihilation operators in the context of fields.

Before we move on, note that the nonzero value $\frac{1}{2}\hbar\omega$ of the ground state energy can be understood in terms of the uncertainty principle: for the energy

$$H = \frac{p^2}{2m} + \tfrac{1}{2}kx^2$$

to vanish would require both $p = 0$ and $x = 0$, but according to uncertainty principles (see p.1046) the more narrowly you fix the location x of the mass m, the less well you know the value of its momentum p and vice versa. The values of p and x in the lowest-energy state are therefore small but nonzero, with the result that the minimal oscillator energy is also nonzero.

22.5.5 Path Integrals

As we saw in the preceding section, it is commutation relations like

$$[p, x] = px - xp = -i\hbar$$

that lead to quantization. This method of quantization is known as canonical quantization. In his doctoral thesis, Feynman formulated an alternative method involving path integrals that is both more powerful and conceptually more interesting. Feynman suggested that the probability amplitude for a particle to go from point A to point B (or for a system to go from state A to state B) was given by

$$\int d\mathcal{P}_{AB}\, e^{iS/\hbar}$$

where S is the action (in the sense of the integral of the Lagrangian that was introduced in Chapter 21) for the particle or system and the $d\mathcal{P}_{AB}$ means that the integral is over all possible paths (trajectories) that take the particle or system from A to B. While path-integral calculations and the proof of the equivalence of path-integral and canonical quantizations are beyond our reach, you can nonetheless see how these path integrals work by noting that

$$e^{iS/\hbar} = \cos\tfrac{S}{\hbar} + i \sin\tfrac{S}{\hbar}$$

For most paths, a slight variation in the path from A to B will correspond to a change in the action S that, though it may be small on an everyday, macroscopic scale, is very large compared to $\hbar$, and the oscillation in $e^{iS/\hbar}$ will therefore be very rapid as the path is varied. The contributions of such paths to the integral over all possible paths will consequently almost completely cancel out by destructive interference: when the contribution from one path is at a peak, that from a closely neighboring path will be at a trough.[26] Only when the action S is at a minimum (and therefore the variation δS in S as paths are varied vanishes) will the contributions to the integral over all possible paths be coherent and reinforce each other. So by far the dominant contribution to the probability amplitude will be from those paths near the path that minimizes the action S — the paths near the trajectory given by the principle of least action in classical (that is, nonquantum) mechanics.

Thus for macroscopic objects with correspondingly macroscopic values of the action S only paths microscopically close to the classical trajectory make a significant contribution to the probability amplitude. In other words, in quantum theory macroscopic objects follow their classical, nonquantum trajectories with effectively dead certainty. For microscopic objects like electrons, however, the action S and its variations δS are more comparable to

[26] For plots showing constructive and destructive interference, see figg. (0.10) and (0.11) on p.46f.

the small scale of $\hbar$, so that even paths that deviate substantially from the classical trajectories make significant contributions to the probability amplitude. An electron headed toward a barrier with two slits, for example, might equally likely pass through either slit, with the result that the combined amplitudes from these two paths create the same wave-like interference pattern of alternating "light" and "dark" bands on the far side of the barrier that one observes for light waves.

If the value of $\hbar$ were macroscopically large — say, 1 Js rather than 10^{-34} Js — , then the same sort of quantum behavior exhibited by electrons and other microscopic particles would be a part of everyday life: when you entered a room through a doorway, for example, instead of your following your classical, straight-line path with effectively dead certainty, inside the room there would be alternating regions of higher and lower probability for where you would end up, some of those regions being quite far from a straight-line path. Kind of cool in some ways, but simple tasks like pouring a cup of coffee, parking a car, shaking hands, or brushing your teeth would become trying ordeals fraught with uncertainty and anxiety. Toothpaste missing the brush and ending up in the sink would get old pretty fast, not to mention having the toothbrush end up in your ear.

22.6 Quantum Field Theory

As in quantum mechanics, in quantum field theory (which also blossomed in the late 1920's) one calculates probability amplitudes. The fundamental quantity in quantum field theory is, however, a field: to each kind of matter, there corresponds a field that permeates all of space and evolves with time, and the value of that field at a particular point in spacetime determines how many particles of that kind of matter there are at that place and time. Electrons and their antiparticles, positrons, for example, are excitations of an electron field, and the excitations of the electromagnetic field are the electromagnetic particle, the photon. That is, if ψ is the electron field, then the value of $\psi(\mathbf{r}, t)$ corresponds to the number of electrons or positrons at location $\mathbf{r}$ at time t. It turns out that these excitations behave mathematically just like a mass on the end of a spring: as was shown in §22.5.4, for a quantized harmonic oscillation, the allowed energies work out (modulo a ground-state energy that need not concern us here) to $E_n = n\hbar\omega$, where ω is the oscillator's natural frequency ($\sqrt{k/m}$ in the case of a literal mass on the end of a spring) and $n = 0, 1, 2, 3, \ldots$. It is the value of n that corresponds to the number of particles, and a quantum field is like an infinite set of harmonic oscillators, one at each point in spacetime.

In quantum mechanics, each of the zillions of electrons in the universe has its own wave function $\psi_{qm}(\mathbf{r}, t)$, with $|\psi_{qm}(\mathbf{r}, t)|^2\, dV$ giving the probability of

finding the electron within an infinitesimal volume dV at location $\mathbf{r}$ at time t. Quantum mechanics simply assumes that all of these zillions of electrons just happen to be identical in charge, mass, and other properties. In quantum field theory, however, there is only one electron field $\psi_{qft}(\mathbf{r}, t)$, with, as just pointed out, the value of $\psi_{qft}(\mathbf{r}, t)$ corresponding to the number of electrons or positrons at location $\mathbf{r}$ at time t. All electrons have the same charge, mass, and other properties in quantum field theory because they are excitations of this single underlying electron field. In this and other respects, quantum field theory is very different from quantum mechanics, both conceptually and in its formulation and predictions.

The mathematical starting point for quantum field theory is a quantity known as the Lagrangian (the same Lagrangian discussed in Chapter 21): once one has written down the Lagrangian, all of the physics follows from it in a unique way according to a technique in the calculus of variations known as the principle of least action.[27] The object of a quantum field theorist is therefore to write down a Lagrangian that accounts for all of the kinds of matter, and all of the kinds of interactions among the kinds of matter, that are observed in our universe.

It turns out that the key to this trick is symmetry: determining the mathematical symmetry of the Lagrangian determines the kinds of matter and interactions that follow from it. Moreover, Lagrangians based on different symmetries can sometimes be combined under the umbrella of a larger symmetry. When this happens, the physics that follows from those various Lagrangians is said to be unified. If one could determine the single, overall symmetry of the universe and write down the corresponding Lagrangian, physics would be complete; all of the physics of our universe — all the kinds of matter that can exist and all of the ways that that matter could interact with itself, everything that could exist or happen physically — could be calculated from this Lagrangian. In effect, all of physics would become just one big homework problem, with the Lagrangian as the given.

The Standard Model is comprised of those aspects of quantum field theory that have so far been abundantly verified experimentally. And by *abundantly* we mean out, in the case of the measurement of the gyromagnetic ratio of the electron, to 11 decimal places — a higher degree of verification than for any other scientific theory in history. The Standard Model has been hugely successful at accounting for most of the kinds of matter and for three of the four forces (interactions) that occur in the universe. Unfortunately, for

[27] For those who have not worked through Chapter 21, while the energy E (more usually known in this context as the Hamiltonian H) is $E = K + U$, the Lagrangian is $L = K - U$. In classical Newtonian mechanics, the principle of least action yields $F = ma$. The great advantage of the Lagrangian formalism is that all of physics can be formulated with it — classical Newtonian mechanics, relativity, quantum mechanics, quantum field theory, string theories, ... — everything.

technical reasons it does not include gravity, and so far it unifies only two of the other three forces.

Among the more noteworthy features of quantum field theory and the Standard Model:

- Everything that exists physically is a quantum field — a field in the mathematical sense of permeating (having a value at every point in) spacetime, and a quantum field in the sense that at any given point in spacetime the field can have only certain discrete values (corresponding to the presence of zero, one, two, . . . , particles of that kind of matter). What we usually think of as particles are the quantum excitations of these fields.

- Some of the fields correspond to what we usually think of as matter, others to what we usually think of as forces (for which the preferable term in this context is interactions). What we usually regard as two particles exerting forces on each other is in fact two particles exchanging other particles of the type corresponding to the force being exerted. For example, in quantum electrodynamics (QED), which is a subset of the Standard Model, there are electrons, positrons, and photons. The electrons and positrons are what we usually think of as the matter. The photons (massless particles of light) turn out to be responsible for the electromagnetic force: electrons and positrons exert electric and magnetic forces on each other by exchanging photons. Thus everything that exists physically is in the form of particles, and everything that happens physically is an exchange of particles.

  That interactions are all just exchanges of particles provides a neat mechanism for conservation of momentum, energy, and other conserved quantities: when a pair of particles interact by exchanging a third particle (known as the vector particle because it conveys the force between the original pair), the momentum and energy lost by the emitting member of the pair is gained by the absorbing member. Newton's third law, that bodies exert equal and opposite forces on each other, is explained by this same mechanism — although at this level in physics no one works with Newton's laws, because "forces" are just an antiquated concept that poorly reflects what's actually happening physically.

- The Lagrangian for the Standard Model, in addition to being relativistic, has certain other simple geometrical symmetries that directly yield the basic conservation laws. For example, invariance of the Lagrangian under spacetime translations (that is, shifts from one place or time to another) yields, as is shown in Appendix B, conservation of momentum and energy: conservation of momentum is a direct consequence of the fact that physics is the same here, there, and everywhere, and conservation of energy is a direct consequence of the fact that physics is the same today as it was yesterday and will

be tomorrow. The actual mathematical proof is marvelously simple and will fit on the back of an envelope (though given in somewhat more expansive detail in Appendix B).

- The kinds of matter that exist and the kinds of interactions that can occur are also determined by symmetries in the Lagrangian. These symmetries, known as gauge symmetries, are more abstract than geometrical symmetries, but, like all symmetries, are characterized by a symmetry operation and a corresponding invariant (that is, conserved quantity). Consider, as an example of geometrical symmetry, a circle: when we say that a circle is rotationally symmetric, we mean that the circle looks the same before and after a rotation. In this case, the symmetry operation is the rotation, and the invariant is the appearance of the circle. An everyday example of a gauge symmetry would be an income indexed to inflation: the symmetry operation would be inflation, and the invariant would be the real value of the income. Having an income indexed to inflation would mean that your income would be adjusted every year so that your spending power remains constant. The gauge symmetries in the Standard Model are of course of a more mathematical nature; their names and the kinds of particles and interactions to which they give rise are

  - $U(1)$ (said "U 1") is the symmetry of quantum electrodynamics (QED) and gives rise to the photon, which conveys the electromagnetic force between electrically charged particles.

    The elements of symmetry group $U(1)$ are the phases $e^{i\varphi}$, which can be identified with the points on the unit circle; the electromagnetic interaction and photons are a direct consequence of the invariance of physics under changes of this phase. In other words, it is as though there were a little circle associated with every point in spacetime, and the electromagnetic interaction is a consequence of physics being independent of one's position on the perimeter of this circle.[28]

  - $SU(2)$ (said "S U 2") is the symmetry of the weak interaction between leptons[29] and quarks, which is conveyed by the $W^{\pm}$ and Z particles. There are three generations of leptons: the electron and positron ($e^{\pm}$), the muons ($\mu^{\pm}$), and the taus ($\tau^{\pm}$), with their corresponding neutrinos and antineutrinos (ν_e, $\bar{\nu}_e$; ν_μ, $\bar{\nu}_\mu$; ν_τ and $\bar{\nu}_\tau$). The weak interaction is so called because its magnitude, compared to that of the electromagnetic and strong interactions, is very weak (see table (22.2) on p.1073). It is of such short range that it is ordinarily confined to nuclei and therefore

[28] The electromagnetic force is explicitly derived from the $U(1)$ symmetry in §22.6.1.

[29] *Lepton* from the Greek for *light* (in the sense of *not heavy*): of all the massive particles, leptons are the lightest.

Quark	Symbol	Charge	Antiquark	Symbol	Charge
Down	d	$-\frac{1}{3}$	Antidown	$\bar{d}$	$+\frac{1}{3}$
Up	u	$+\frac{2}{3}$	Antiup	$\bar{u}$	$-\frac{2}{3}$
Strange	s	$-\frac{1}{3}$	Antistrange	$\bar{s}$	$+\frac{1}{3}$
Charm	c	$+\frac{2}{3}$	Anticharm	$\bar{c}$	$-\frac{2}{3}$
Bottom	b	$-\frac{1}{3}$	Antibottom	$\bar{b}$	$+\frac{1}{3}$
Top	t	$+\frac{2}{3}$	Antitop	$\bar{t}$	$-\frac{2}{3}$

Table 22.1: The Quarks

not directly familiar to you from daily experience, but it is, for example, the mechanism for nuclear beta decay.

– $SU(3)$ (said "S U 3") is the symmetry of the strong interaction conveyed between quarks by gluons. Combinations of quarks bound together by gluons make up the baryons (protons, neutrons, and other similar particles) and the mesons (pions ($\pi^{\pm}$, π^0), etc.).[30] Corresponding to the 3 in $SU(3)$, quarks come in three colors — though the term *color* is here used in a whimsical and abstract sense that has nothing to do with the colors of the visual spectrum. Quarks also come in various flavors, as shown, from the lightest mass in the top row to the heaviest in the bottom row, in table (22.1).[31] Each flavor consists of a quark and antiquark pair, with the quark carrying an electric charge, in units of the fundamental charge $e = 1.6 \times 10^{-19}$ C, of either $-\frac{1}{3}$ or $+\frac{2}{3}$, and the corresponding antiquark having the same mass but opposite charge. The pairs or doublets up-down, charm-strange, and top-bottom constitute the three generations of quarks corresponding to the three generations of leptons.[32] The protons and neutrons of which nuclei are composed are combinations of up and down quarks: up-up-down and up-down-down, respectively. The strong interaction is so called because, although it is confined to distance scales of nuclear size, it is much stronger than

[30] Baryons and mesons are so called because (again from the Greek) they are the heaviest and the intermediate among the observed massive particles. Together, baryons and mesons constitute the hadrons (from the Greek for *pleasantly plump*).

[31] The bottom and top quarks at first also went by the competing names *beauty* and *truth*, but unfortunately the prosaic ultimately triumphed over the poetic.

[32] While quantum field theoretical reasons limit the number of quark flavors to 16 (eight generations) and also require that the number of generations of quarks and leptons be equal, there are reasons for believing that the three generations we have observed are all there are: experimental results for the Z lifetime are consistent with three and only three generations, and in string theory certain promising Calabi-Yau manifolds (the mathematical terrain on which the strings live) give rise to three generations.

the electromagnetic and weak interactions (see table (22.2) on p.1073). This relative strength is in fact how the strong interaction overcomes the electromagnetic repulsion between positively charged protons and holds the nucleus together. Its physics is known — by analogy to quantum electrodynamics, with *electro-* replaced by *chromo-* for color — as quantum chromodynamics (QCD).

- Local Poincaré invariance (invariance under local Lorentz transforms, rotations, and translations) is the symmetry responsible for the gravitational interaction, which is conveyed by gravitons. All particles undergo gravitational interactions, even massless particles (since their energy is equivalent to a mass according to $E = mc^2$). For technical mathematical reasons (about which more below), gravity is not included in the Standard Model. Gravitons have not yet been observed, but we already know their properties (massless, spin 2, etc.).

- Of the three symmetries $U(1)$, $SU(2)$, and $SU(3)$, the $U(1)$ of electromagnetism and the $SU(2)$ of the weak interaction have been unified by combining them into a higher symmetry. At high energies (like those in the early universe), this higher symmetry is obeyed, and there is no difference between the electromagnetic and weak interactions; there is only a single electroweak interaction. But at lower energies, this higher symmetry is violated by a mechanism known as spontaneous symmetry breaking and the electromagnetic and weak interactions are distinct. Without going into the mathematical details, the basic idea is this: the Lagrangian, and therefore the physics of the universe, actually always obeys the full, higher symmetry, but the vacuum (everything around us, less the particles) does not: as the energy density of the universe drops, the vacuum, as it were, freezes in some random "direction" that breaks the symmetry.

  Though it is much more abstract mathematically, spontaneous symmetry breaking in quantum field theory is very similar to what happens magnetically to a lump of iron. The physical theory of electromagnetism is symmetric; there is no preferred or special direction for magnetic fields. At high energies (high temperatures), the magnetic fields of the iron atoms are being jostled too much by random thermal motion to align in any particular direction, so that the state of the lump of iron (analogous to the vacuum in quantum field theory) is also symmetric. But as the temperature is lowered, adjacent iron atoms will tend to align magnetically with each other more and more, until the lump of iron ends up with all of its atoms in alignment.[33] This

[33] Actually, this process begins independently at various sites in the lump, with the result that the lump is divided into domains, the field being uniform in direction within each domain but differing in direction between domains. But what we are saying is true within each domain. No blood, no foul.

alignment defines an axis, a special direction, that violates the symmetry of electromagnetism. This direction of alignment is random and arises spontaneously as the iron cools. And it is only the lump of iron that has broken the symmetry: the laws governing electromagnetism remain symmetric.

Particles that convey interactions associated with unbroken symmetries are massless, but, by what is known as the Higgs mechanism, the spontaneous breaking of the higher symmetry that unifies the weak and electromagnetic interactions gives mass to the $W^{\pm}$ and Z particles that convey the weak interaction and also involves a new kind of particle called the Higgs particle. Electroweak unification successfully predicts the masses of the $W^{\pm}$ and Z, and while the Higgs was not seen for many years, this was not surprising, as its interactions were expected to be difficult to discern. It was not until the increase in energy at the Large Hadron Collider that the Higgs was conclusively produced.[34]

- Three of the more basic symmetries of the Lagrangian and thus of the universe are charge conjugation (C), parity (P), and time reversal (T). Charge conjugation interchanges all particles with their antiparticles. Parity pulls space inside out, so that left is interchanged with right.[35] Time reversal, as you might expect, runs time backward. The combination CPT — what you get by interchanging particles with antiparticles, pulling space inside out, and reversing time, all together — is known to be good for all of the four interactions in nature; indeed, the entire mathematical framework of quantum field theory would fall apart if it weren't, and it is hard to imagine any way to formulate physics in a universe that did not have this symmetry.[36] Three of the four interactions also obey C, P, and T individually, as well as all combinations of them (like CP, etc.), but the weak interaction has been found to violate all but the full combination CPT.[37]

[34]For the papers announcing the discovery, see `https://arxiv.org/abs/1207.7235` and `https://arxiv.org/abs/1207.7214`.

[35]To visualize space being pulled inside out, you can think of pulling a glove inside out: a glove that fits your right hand will, when pulled inside out, fit your left hand. (Provided, of course, that the glove has a distinct front and back — the latex gloves in the lab do not, so you have to use your imagination or mark one side if you are going to use them to demonstrate this effect.)

[36]Recently there apparently have been some indications that CPT may be violated in quantum theories of gravity, as well as in some other radical, left-wing theories, but we haven't had time to read up on those because we've had to spend so much time writing this nonsense.

[37]The weak interaction in fact violates parity maximally: only left-handed electrons are involved in beta decay and other weak interactions. (You are probably used to thinking of the spin on electrons as being "up" or "down" — which is fine as long as you have some axis by which to define up and down. In the absence of such an externally defined axis, the electron's spin is measured along the axis of its own momentum: if, when you point the thumb of your right hand in the direction of the electron's momentum, your

Strong	10
Electromagnetic	$\frac{1}{137}$
Weak	10^{-7}
Gravitational	10^{-45}

Table 22.2: Approximate Relative Strengths of the Interactions

- Of the four interactions, we are familiar with the electromagnetic and gravitational on our everyday, macroscopic scale because they are long-range; the weak and strong interactions are less familiar because they are of very short range, on the order of the size of nucleus or smaller. The relative strengths of the four interactions are roughly as shown in table (22.2). Note that gravity is the weakest of the four interactions in nature by about 40 orders of magnitude; we experience substantial gravitational effects only because *all* particles interact gravitationally and because that gravitational interaction is always an attraction, with the result that net gravitational effects can, cumulatively, be significant. In contrast, the electromagnetic interaction, while it acts between all of the myriad electrically charged particles in ordinary matter, is generally very small on a macroscopic scale because the objects around us typically carry little or no net charge and consequently the attractions of unlike charges and the repulsions of like charges largely or completely cancel each other out.

- The strong interaction turns out to get stronger as the distance between quarks increases, so much so that it is impossible to separate out individual quarks; they always come in at least pairs. If you tried to pull apart a down–anti-down pair ($d\bar{d}$), for example, you would have to put more and more energy into pulling the pair farther and farther apart, and at some point you would have put in so much energy that another quark-antiquark pair would be created — so that you would end up, not with a d and a $\bar{d}$, but with something like two $d\bar{d}$ pairs. This impossibility of separating out individual quarks is known as confinement.

  Thus it turns out that although the quarks have fractional charges ($\pm\frac{1}{3}$ or $\pm\frac{2}{3}$), because of color confinement they occur only in combinations that have whole multiples of the fundamental charge. A proton, for example, consists of two up quarks and one down quark, so that it's charge is $\frac{2}{3}+\frac{2}{3}-\frac{1}{3}=1$.

---

fingers curl around in the direction of its spin, the electron is said to be right-handed; otherwise, it is left-handed. This "handedness" is known as chirality (from the Greek for "handedness"). Right- and left-handed spins are sometimes also referred to as positive and negative helicity.)

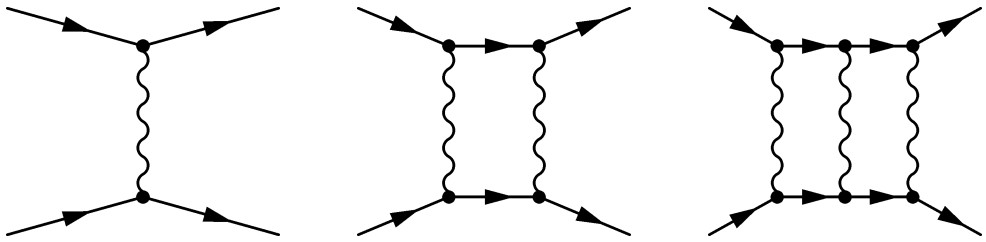


Figure 22.1: Feynman Diagrams

- Although in principle the equations generated by the Lagrangian for the Standard Model could be solved exactly in closed form, the math to work out these solutions doesn't yet exist. Instead, calculations have to be done by means of perturbation theory: it turns out that often the interactions between particles are rather small corrections to the solutions for free (that is, noninteracting) particles, so that one can take the solutions for free particles as a lowest-order solution and make a series of successive approximations, allowing for ever more complicated interactions. This method has been tremendously successful, but it has limitations (see the following item).

  In this scheme, the interactions that particles undergo turn out to be expressed mathematically by large and complicated integrals. These interactions can, however, also be represented by simple pictures known, after their inventor, as Feynman diagrams. The particles themselves are represented by lines, and their interactions are the vertices where these lines intersect. Three of the simplest contributions to the interaction between a pair of electrons, for example, come from the "ladder" diagrams, shown in fig. (22.1), in which one or more photons are directly exchanged.[38] Here the solid lines represent the electrons, the wavy lines the exchanged photons. The arrows indicate the direction of motion of the electrons; arrows have been omitted from the photon lines because the photons could be going in either direction. Feynman diagrams are a simple, visual way of understanding how particles interact, and each diagram corresponds to a precise mathematical contribution to the probability amplitude for an interaction.

  In principle, to solve for the scattering of the two electrons off of each other, you need to sum up the contributions to the interaction between them from

[38] Classes of Feynman diagrams tend to be named according to what they look like in everyday terms. In addition to ladder diagrams, there are tadpoles, amputated diagrams, tree-level diagrams, triangle diagrams, and a host of others. There is even a penguin diagram, which purportedly got its name when John Ellis, as a result of losing a dart game in a pub, was required to use the word *penguin* in his next paper. When neatly drafted in a typeset publication, the diagram doesn't look all that much like a penguin, but the name has stuck.

each of the infinitely many possible Feynman diagrams.[39] In practice, the more complicated the diagram, the smaller the contribution, so that you can get a good approximation by working out a relatively small number of the simpler diagrams, depending on how accurate an answer you need. Note that there are closed loops in the two rightmost diagrams in fig. (22.1): one way of measuring the complexity of diagrams is by the number of closed loops. One thus speaks of calculating a scattering amplitude "to two loops" and so on.[40]

- People worked on further developing quantum field theory after its beginnings in the late 1920's, but they quickly ran into a major difficulty that it took about twenty years to overcome. The problem is that each closed loop in a Feynman diagram corresponds to an integration over a momentum that ranges over all possible values, from $-\infty$ to $+\infty$. This infinite range of integration causes the resulting probability amplitude to diverge — which, since probabilities are supposed to come out between 0 and 1 (0 and 100%), presents a problem.

  The solution to this difficulty is known as renormalization: it turns out to be possible mathematically to "absorb" these problematic infinities into the particle masses and coupling constants (the parameters that measure the strengths of the various interactions). In a sense, renormalization amounts to the observation that the sum of all the positive integers is $-\frac{1}{12}$:[41]

$$1 + 2 + 3 + 4 + \cdots = -\frac{1}{12}$$

  Carrying out renormalization involves temporarily working in $4+\epsilon$ spacetime dimensions, where ϵ is a complex number, and then, when the calculation is complete, taking the limit as $\epsilon \to 0$. Factors that diverge in this limit (as well as some that don't) are disposed of by supposing that other factors conveniently go to zero in a way that exactly compensates for the divergence. The whole business is enough to make you wonder whether physicists have any principles or integrity at all. But, lo and behold, at the end of all these shenanigans one is left with predictions for the values of observable quantities that agree spectacularly well with experiment.

  For three of the four interactions — electromagnetic, weak, and strong — the degree of divergence (the "badness" or power of the unwanted infinity)

[39] Actually, this is fibbing a little: the series expansion of perturbation theory is asymptotic, not convergent.

[40] The number of diagrams increases exponentially with the number of loops. Toichiro Kinoshita at Cornell has spent much of his life doing a perturbative calculation of the anomalous magnetic moments of the electron and muon to ever greater accuracy. The one-loop and two-loop calculations involve just a few diagrams, but by three loops there are 72 diagrams, and it took Kinoshita three years to tackle the 891 diagrams at four loops. Kinoshita is presently working on the 12,672 five-loop diagrams.

[41] For the skeptical, this is proved in Appendix E.

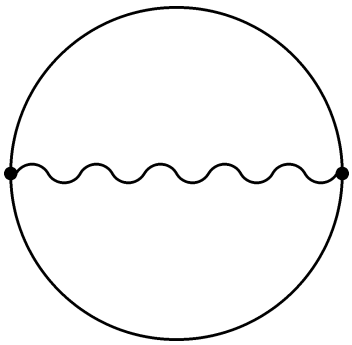


Figure 22.2: A Vacuum Bubble

remains constant regardless of the number of loops, and renormalization is therefore possible. Unfortunately, the degree of divergence of the gravitational interaction gets worse the greater the number of loops, so it can't be renormalized. This does not mean that the quantum field theory of gravity is a bad theory; it just means that perturbation theory, the only mathematical means we have of working out quantum field theories, doesn't work for gravity. This is just one of many technical mathematical difficulties that at present prevent us from formulating a quantum theory of gravity.

- In quantum field theory, the vacuum (empty space) turns out to be seething with all sorts of activity. In classical physics, you can't get something from nothing, but in quantum field theory, there are Feynman diagrams that correspond to doing exactly that. The Feynman diagram in fig. (22.2) shows an electron-positron pair and a photon spontaneously coming into existence out of nothing and subsequently annihilating each other to return to nothing. It turns out that virtual particles — particles corresponding to closed loops in Feynman diagrams — , do not, as the term goes, have to be on the mass shell: virtual particles, unlike real (that is, observable) particles, can have energies and momenta that violate

$$E^2 - p^2c^2 = m^2c^4$$

with the result that they can exhibit this kind of odd behavior. The vacuum is actually full of these sorts of vacuum bubbles. But while vacuum bubbles occur all the time, they are, since no observable particles go into or come out of them, intrinsically unobservable (unless of course you are one of the particles *in* the bubble — which has consequences for the origin of the universe that we'll discuss in §22.8).

- While quantum field theory and the Standard Model account extremely well for much of the physical universe, it has limitations and defects:
  - No one has been able to figure out a way to quantize gravity, so it cannot be included in the Standard Model.
  - Although the electromagnetic and weak interactions have been unified, no one has yet found a scheme that successfully unifies this combined electroweak interaction with the strong interaction.
  - In addition to the coupling constant for gravity, which it excludes, the Standard Model involves 19 arbitrary parameters[42] and its Lagrangian is a correspondingly sprawling mess:[43]

$$\begin{aligned}
\mathcal{L} = &-\tfrac{1}{4}(\partial_\mu G^i_\nu - \partial_\nu G^i_\mu - g_3 f^{ijk} G^j_\mu G^k_\nu)^2 \\
&- \tfrac{1}{4}(\partial_\mu W^a_\nu - \partial_\nu W^a_\mu - g_2 \epsilon^{abc} W^b_\mu W^c_\nu)^2 \\
&- \tfrac{1}{4}(\partial_\mu B_\nu - \partial_\nu B_\mu)^2 \\
&- \sum_n \bar{q}_{nL} \gamma^\mu (\partial_\mu + \tfrac{i}{2} g_3 \lambda^i G^i_\mu + \tfrac{i}{2} g_2 \tau^a W^a_\mu + \tfrac{i}{6} g_1 B_\mu) q_{nL} \\
&- \sum_n \bar{u}_{nR} \gamma^\mu (\partial_\mu + \tfrac{i}{2} g_3 \lambda^i G^i_\mu + \tfrac{2i}{3} g_1 B_\mu) u_{nR} \\
&- \sum_n \bar{d}_{nR} \gamma^\mu (\partial_\mu + \tfrac{i}{2} g_3 \lambda^i G^i_\mu - \tfrac{i}{3} g_1 B_\mu) d_{nR} \\
&- \sum_n \bar{\ell}_{nL} \gamma^\mu (\partial_\mu + \tfrac{i}{2} g_2 \tau^a W^a_\mu - \tfrac{i}{2} g_1 B_\mu) \ell_{nL} \\
&- \sum_n \bar{e}_{nR} \gamma^\mu (\partial_\mu - i g_1 B_\mu) e_{nR} \\
&+ \left| (\partial_\mu + \tfrac{i}{2} g_2 \tau^a W^a_\mu - \tfrac{i}{2} g_1 B_\mu) \phi \right|^2 + \mu^2 \phi^\dagger \phi - \lambda (\phi^\dagger \phi)^2 \\
&- \sum_{m,n} (\Gamma^u_{mn}\, \bar{q}_{mL}\, \phi^c\, u_{nR} + \Gamma^d_{mn}\, \bar{q}_{mL}\, \phi\, d_{nR} + \Gamma^e_{mn}\, \bar{\ell}_{mL}\, \phi\, e_{nR} + \text{h.c.})
\end{aligned} \tag{22.23}$$

  Ideally, a truly fundamental theory of nature would have at most one parameter — for the strength of its single, fully unified interaction — and the values of all other parameters would be determined by the theory.

[42] If you were curious: 3 coupling constants, one each for the strong, weak, and electromagnetic interactions; 6 quark masses; 3 generalized Cabibbo angles and 1 CP-violating phase angle (which have to do with the weak interaction); 2 parameters related to the Higgs field; 3 lepton masses; and 1 vacuum phase angle (from nonperturbative effects in quantum chromodynamics). If a nonzero mass is included for neutrinos, there will be several more.

[43] Adapted from the much more industrious G. Rajasekaran, *From Atoms to Quarks and Beyond: A Historical Panorama* (`https://arxiv.org/abs/physics/0602131`).

22.6.1 The Electromagnetic Force from Symmetry

This is quite a reach from where we are now, but the way in which forces arise from symmetries is so beautiful and such a fundamentally important mechanism in physics that we have to try at least to give you the basic idea.

Our derivation will be in four steps:

Step 1 The Lagrangian for free (that is, noninteracting) electrons and positrons will be introduced. We will show that the corresponding equation of motion, known as the Dirac equation, leads to electrons and positrons that satisfy the relativistic relation $E^2 - p^2c^2 = m^2c^4$.[44] Unfortunately we won't be able to do much more with the Dirac equation than that — its origins and other properties will have to remain obscure — , but that shouldn't be a major bummer because our focus is rather on showing how the electromagnetic force arises; the Dirac equation will simply be our assumed starting point. Just don't let the mysteriousness and apparent complexity of this step freak you out and discourage you from continuing.

Step 2 We will show that imposing a $U(1)$ symmetry on the electron field requires the introduction of a new field, called the gauge field.

Step 3 We will show that the corresponding symmetry of this gauge field, which is known as a gauge symmetry, is the gauge symmetry of the Maxwell equations.

Step 4 We will show that the most general allowed contribution of this new gauge field to the Lagrangian results in an equation of motion that reproduces the Maxwell equations and hence the electromagnetic force.

First, a couple of preliminaries.[45] Up to now, we have been using the system of electromagnetic units favored by engineers, which involves $4\pi\epsilon_0$ and μ_0. No self-respecting physicist uses this system of units, certainly not when doing field theory. Moreover, it is for convenience customary when doing quantum field theory to work also in natural units, where the mass, length, and time scales are taken to be such that $\hbar = 1$ and $c = 1$.[46] If

[44] This demonstration is actually not essential to the derivation of the electromagnetic force from symmetry; it is merely an attempt to do at least something to make the Dirac equation and the Lagrangian that leads to it more plausible.

[45] Not that preliminaries could come other than first.

[46] For $c = 1$, this would mean, for example, measuring lengths in meters and time in increments corresponding to how long it takes light to travel 1 m: $1\,\text{m}/3.0 \times 10^8\,\text{m/s} = 3.3 \times 10^{-9}\,\text{s}$. If we call this time unit, say, a 'tu,' then, since light travels one 1 m in 1 tu, the speed of light is $c = 1\,\text{m}/1\,\text{tu} = 1\,\text{m/tu}$. Setting $\hbar = 1$ would mean defining a mass unit in a similar manner.

we work in natural units and a system of electromagnetic units that doesn't involve the $4\pi\epsilon_0$ and μ_0, then the Maxwell equations take the simpler form

$$\nabla \cdot \mathbf{E} = \rho \tag{22.24a}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{22.24b}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{22.24c}$$

$$\nabla \times \mathbf{B} = \mathbf{j} + \frac{\partial \mathbf{E}}{\partial t} \tag{22.24d}$$

Also, relativistic relations that would have involved ct now, since $c = 1$, involve simply t. Relations like $E^2 - |\mathbf{p}|^2 c^2 = m^2 c^4$ will become simply $E^2 - |\mathbf{p}|^2 = m^2$, etc.

We also need to introduce you to index notation. The big-people way to write the four-vector (ct, x, y, z) is x^μ, where the index μ runs from 0 to 3, with x^0 corresponding to ct and x^1, x^2, and x^3 to x, y, z.[47] We will also use what is known as the summation convention: in terms where an index is repeated, that index is understood to be summed over. Thus, for example, $\partial x^\mu / \partial x^\mu$ stands for

$$\frac{\partial x^\mu}{\partial x^\mu} = \sum_{\mu=0}^{3} \partial_\mu x^\mu = \frac{\partial (ct)}{\partial (ct)} + \frac{\partial x}{\partial x} + \frac{\partial y}{\partial y} + \frac{\partial z}{\partial z} = 1 + 1 + 1 + 1 = 4$$

In special relativity, the square ds^2 of the invariant interval,

$$\begin{aligned} ds^2 &= (c\,dt)^2 - dx^2 - dy^2 - dz^2 \\ &= (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2 \end{aligned}$$

can be neatly written as

$$ds^2 = \eta_{\mu\nu}\, dx^\mu\, dx^\nu \tag{22.25}$$

where $\eta_{\mu\nu}$ is the matrix, known as the flat-space metric, defined as

$$\eta_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

To see this, we have only to do out the multiplication in eq. (22.25):

$$\eta_{\mu\nu}\, dx^\mu\, dx^\nu = dx^\mu\, \eta_{\mu\nu}\, dx^\nu$$

[47] Of course, in natural units the ct's here will be simply t's, but for the time being we are keeping the c's to try to make the connection with our earlier work with four-vectors more clear.

$$
\begin{aligned}
&= \sum_{\mu=0}^{3}\sum_{\nu=0}^{3} dx^\mu\, \eta_{\mu\nu}\, dx^\nu \\
&= \begin{pmatrix} dx^0 & dx^1 & dx^2 & dx^3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} dx^0 \\ dx^1 \\ dx^2 \\ dx^3 \end{pmatrix} \\
&= \begin{pmatrix} dx^0 & dx^1 & dx^2 & dx^3 \end{pmatrix} \begin{pmatrix} dx^0 & 0 & 0 & 0 \\ 0 & -dx^1 & 0 & 0 \\ 0 & 0 & -dx^2 & 0 \\ 0 & 0 & 0 & -dx^3 \end{pmatrix} \\
&= (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2 \\
&= ds^2 \quad \checkmark
\end{aligned}
$$

Superscript indices are called contravariant and subscript indices covariant. Note that, as we have just seen, Lorentz-invariant quantities like ds^2 always involve sums over pairs of indices, one of which is contravariant and the other covariant. The sum over such pairs of indices is called contraction and is the only sensible way to sum over a pair of indices: the contraction of two contravariant or of two covariant indices would yield quantities that are not Lorentz invariant and that are therefore physically gibberish. The contraction of two contravariant indices would, of example, yield

$$
\begin{aligned}
dx^\mu dx^\mu &= (dx^0)^2 + (dx^1)^2 + (dx^2)^2 + (dx^3)^2 \\
&= c^2\, dt^2 + dx^2 + dy^2 + dz^2
\end{aligned}
$$

— a combination that transforms in a bizarre way between reference frames. One can in fact think of $\eta_{\mu\nu}$ as an animal that lowers indices: since

$$
\eta_{\mu\nu}\, dx^\nu = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} dx^0 \\ dx^1 \\ dx^2 \\ dx^3 \end{pmatrix} = \begin{pmatrix} dx^0 \\ -dx^1 \\ -dx^2 \\ -dx^3 \end{pmatrix}
$$

we can write

$$
ds^2 = dx^\mu dx_\mu
$$

where

$$
\begin{aligned}
dx^\mu &= (dx^0, dx^1, dx^2, dx^3) \\
dx_\mu &= \eta_{\mu\nu}\, dx^\nu = (dx^0, -dx^1, -dx^2, -dx^3)
\end{aligned}
\tag{22.26}
$$

So much for preliminaries. Now down to business.

Step 1: The electron-positron field is written $\psi(x)$, where ψ itself is a four-component column vector (known as a Dirac spinor) and the x in $\psi(x)$ is a shorthand notation for the spacetime point x^μ, that is, for (t, x, y, z).[48] The Lagrangian density (Lagrangian per unit volume) for ψ is

$$\mathcal{L} = \overline{\psi}(x)(i\not{\partial} - m)\psi(x) \tag{22.27}$$

You probably find that this requires some explanation. Here m is the electron mass. The slash on the ∂ is called a Feynman slash and is a shorthand notation for

$$\not{\partial} = \gamma^\mu \partial_\mu = \gamma^\mu \frac{\partial}{\partial x^\mu}$$

where the γ^μ are a set of four matrices (γ^0, γ^1, γ^2, and γ^3), each of which is 4×4. The exact nature of these gamma matrices need not concern us, beyond the fact that they have the curious property that

$$\gamma_\mu \gamma_\nu + \gamma_\nu \gamma_\mu = 2\eta_{\mu\nu} \tag{22.28}$$

The bar over the ψ on the left is also a shorthand notation: it stands for

$$\overline{\psi} = \psi^\dagger \gamma^0$$

where the $\dagger$ indicates the Hermitian conjugate (that is, the complex conjugate of the transpose).

If all of this is freaking you out, try breathing into a paper bag until you calm down.

There. That's better.

The Lagrange equation of motion for $\overline{\psi}$ is actually pretty simple:

$$\frac{\partial}{\partial x^\mu} \frac{\partial \mathcal{L}}{\partial\left(\frac{\partial \overline{\psi}}{\partial x^\mu}\right)} - \frac{\partial \mathcal{L}}{\partial \overline{\psi}} = 0$$

Since $\mathcal{L}$ doesn't depend on derivatives of ψ, the first term vanishes, and the derivative in the second term just snarfs up the $\overline{\psi}$ in $\mathcal{L}$, so that the equation of motion works out to simply

$$(i\not{\partial} - m)\psi = 0 \tag{22.29}$$

Eq. (22.29) is the Dirac equation for a free electron field. If we multiply both sides of eq. (22.29) by $(i\not{\partial} + m)$, the cross-terms will cancel and we will be left with

$$\left((i\not{\partial})^2 - m^2\right)\psi = 0 \tag{22.30}$$

[48] Henceforward we are using natural units.

Now, from eq. (22.8) on p.1052 we see that the generalization of the quantum momentum operator to the relativistic case will be [49]

$$p_\mu = -i\frac{\partial}{\partial x^\mu}$$

so that

$$(i\not{\partial})^2 = (-i\not{\partial})^2 = \left(-i\gamma^\mu\frac{\partial}{\partial x^\mu}\right)\left(-i\gamma^\nu\frac{\partial}{\partial x^\nu}\right) = (\gamma^\mu p_\mu)(\gamma^\nu p_\nu) = p_\mu p_\nu \gamma^\mu\gamma^\nu$$

Eq. (22.30) thus becomes

$$(p_\mu p_\nu \gamma^\mu\gamma^\nu - m^2)\psi = 0 \tag{22.31}$$

Next, note that it doesn't matter how an index that is summed over is labeled, in the sense that, for example,

$$\sum_{n=0}^{3} x_n^2 = \sum_{m=0}^{3} x_m^2$$

Such repeated indices are called dummy indices, and they may therefore be relabeled with the same freedom that dummy variables can be relabeled in integrations. So there is no reason why we can't relabel the μ and ν in $p_\mu p_\nu \gamma^\mu\gamma^\nu$ by interchanging them:

$$p_\mu p_\nu \gamma^\mu\gamma^\nu = p_\nu p_\mu \gamma^\nu\gamma^\mu \tag{22.32}$$

This relabeling is not as trivial as it might seem, because although the momenta p_μ and p_ν commute with each other, the matrices γ^μ and γ^ν do not. Because of eq. (22.32), we have

$$\begin{aligned}\tfrac{1}{2}p_\mu p_\nu(\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu) &= \tfrac{1}{2}p_\mu p_\nu\gamma^\mu\gamma^\nu + \tfrac{1}{2}p_\nu p_\mu\gamma^\nu\gamma^\mu \\ &= \tfrac{1}{2}p_\mu p_\nu\gamma^\mu\gamma^\nu + \tfrac{1}{2}p_\mu p_\nu\gamma^\mu\gamma^\nu \\ &= p_\mu p_\nu\gamma^\mu\gamma^\nu\end{aligned}$$

which, in conjunction with $\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu = 2\eta^{\mu\nu}$ from eq. (22.28), gives

$$p_\mu p_\nu \gamma^\mu\gamma^\nu = \tfrac{1}{2}p_\mu p_\nu(\gamma^\mu\gamma^\nu + \gamma^\nu\gamma^\mu) = \tfrac{1}{2}p_\mu p_\nu(2\eta^{\mu\nu}) = p_\mu p_\nu\eta^{\mu\nu}$$

Finally, by virtue of eq. (22.26) we arrive at

$$\begin{aligned}p_\mu p_\nu\gamma^\mu\gamma^\nu &= p_\mu p_\nu\eta^{\mu\nu} \\ &= p_\mu p^\mu\end{aligned}$$

[49] Remember that in natural units $\hbar = 1$.

$$= \begin{pmatrix} p^0 & -p^1 & -p^2 & -p^3 \end{pmatrix} \begin{pmatrix} p^0 \\ p^1 \\ p^2 \\ p^3 \end{pmatrix}$$
$$= (p^0)^2 - (p^1)^2 - (p^2)^2 - (p^3)^2$$

To revert to our more usual three-vector notation, recall, from §10.7, that the time component p^0 of the momentum was the energy E, and that

$$(p^1)^2 + (p^2)^2 + (p^3)^2 = p_x^2 + p_y^2 + p_z^2 = |\mathbf{p}|^2$$

so that

$$p_\mu p_\nu \gamma^\mu \gamma^\nu = E^2 - |\mathbf{p}|^2$$

Thus eq. (22.31) becomes

$$(E^2 - |\mathbf{p}|^2 - m^2)\psi = 0$$

which means that for the electron field

$$E^2 - |\mathbf{p}|^2 - m^2 = 0$$

This is just the usual relativistic relation among mass, energy, and momentum, $E^2 - |\mathbf{p}|^2 = m^2$.

Step 2: The elements of the symmetry group $U(1)$ are just the points on the unit circle, which we can specify by $e^{i\theta}$. We now impose on the electron field a $U(1)$ symmetry, that is, we suppose that the transform

$$\psi \to \psi' = e^{i\theta}\psi \tag{22.33}$$

has no effect on the resulting physics. Since the Lagrangian determines all of the physics, to impose this symmetry we need only ensure that the transform (22.33) leaves the Lagrangian invariant. Recalling that the † operation involves taking the complex conjugate and that $\overline{\psi} = \psi^\dagger \gamma^0$, we have

$$\begin{aligned} \mathcal{L}' &= \overline{\psi'}(x)(i\partial\!\!\!/ - m)\psi'(x) \\ &= (\psi'(x)\gamma^0)^\dagger (i\partial\!\!\!/ - m)\psi'(x) \\ &= (e^{i\theta}\psi(x)\gamma^0)^\dagger (i\partial\!\!\!/ - m)e^{i\theta}\psi(x) \\ &= e^{-i\theta}(\psi(x)\gamma^0)^\dagger (i\partial\!\!\!/ - m)e^{i\theta}\psi(x) \\ &= e^{-i\theta}\,\overline{\psi}(x)(i\partial\!\!\!/ - m)e^{i\theta}\psi(x) \end{aligned} \tag{22.34}$$

The $e^{-i\theta}$ will thus cancel against $e^{i\theta}$, leaving us with the original Lagrangian:

$$\mathcal{L}' = \overline{\psi}(x)(i\partial\!\!\!/ - m)\psi = \mathcal{L}$$

At least, this will be true when θ is a constant. Since a θ that is constant is independent of spacetime location x and therefore has the same effect on $\psi(x)$ at all spacetime locations, transforms of the form (22.33) with constant θ are called a global transforms. But suppose now that we try to make the symmetry of the electron field under these transforms local, so that the value of θ can vary from one spacetime location to another, that is, so that θ is now a function $\theta = \theta(x)$ of spacetime location x. In eq. (22.33), this presents us with the complication that the derivative $\partial\!\!\!/$ will now act not only on ψ but also on the $e^{i\theta}$, which leaves us with an extra term:

$$\begin{aligned}
\mathcal{L}' &= e^{-i\theta(x)}\,\overline{\psi}(x)(i\partial\!\!\!/ - m)e^{i\theta(x)}\psi(x) \\
&= e^{-i\theta(x)}\,\overline{\psi}(x)\left(i\gamma^\mu\frac{\partial}{\partial x^\mu} - m\right)e^{i\theta(x)}\psi(x) \\
&= e^{-i\theta(x)}\,\overline{\psi}(x)\left(i\gamma^\mu\frac{\partial e^{i\theta(x)}}{\partial x^\mu}\psi(x) + i\gamma^\mu e^{i\theta(x)}\frac{\partial\psi(x)}{\partial x^\mu} - me^{i\theta(x)}\psi(x)\right) \\
&= e^{-i\theta(x)}\,\overline{\psi}(x)\left(i\gamma^\mu e^{i\theta(x)}i\frac{\partial\theta(x)}{\partial x^\mu}\psi(x) + i\gamma^\mu e^{i\theta(x)}\frac{\partial\psi(x)}{\partial x^\mu} - me^{i\theta(x)}\psi(x)\right)
\end{aligned}$$

which, if we now cancel the $e^{i\theta(x)}$'s against the $e^{-i\theta(x)}$ and set $i^2 = -1$ in the first term, reduces to

$$\begin{aligned}
&= \overline{\psi}(x)\left(-\gamma^\mu\frac{\partial\theta(x)}{\partial x^\mu}\psi(x) + i\gamma^\mu\frac{\partial\psi(x)}{\partial x^\mu} - m\psi(x)\right) \\
&= \overline{\psi}(x)\left(-\gamma^\mu\frac{\partial\theta(x)}{\partial x^\mu}\psi(x) + i\partial\!\!\!/\psi(x) - m\psi(x)\right) \\
&= \overline{\psi}(x)\left(-\gamma^\mu\frac{\partial\theta(x)}{\partial x^\mu} + i\partial\!\!\!/ - m\right)\psi(x)
\end{aligned}$$

This extra term involving the derivative of θ is preventing us from reducing the right-hand side to simply $\mathcal{L}$, thereby messing up the symmetry. To restore the symmetry, we have to go back to the original Lagrangian and include a new term $eA\!\!\!/ = e\gamma^\mu A_\mu$ between the $\overline{\psi}$ and the ψ, where $-e$ is the electron's electric charge and A_μ is some new field, the properties of which remain to be determined. (Though it may seem odd for us to have written this new term as $eA\!\!\!/$ rather than simply $A\!\!\!/$, since we are the ones defining A_μ, we can pull any factor we want out of its definition, and pulling out a factor of e will later prove convenient.) With this new term, we have

$$\mathcal{L} = \overline{\psi}(x)(i\partial\!\!\!/ + eA\!\!\!/ - m)\psi(x) \tag{22.35}$$

and

$$\mathcal{L}' = \overline{\psi}(x)\left(-\gamma^\mu\frac{\partial\theta(x)}{\partial x^\mu} + i\partial\!\!\!/ + eA\!\!\!/ - m\right)\psi(x)$$

$$= \overline{\psi}(x)\left(-\gamma^\mu \frac{\partial\theta(x)}{\partial x^\mu} + i\not{\partial} + e\gamma^\mu A_\mu - m\right)\psi(x) \tag{22.36}$$

Now, if we suppose that the effect of the $U(1)$ transform is not only to take, according to eq. (22.33),

$$\psi \rightarrow \psi' = e^{i\theta}\psi$$

but also, at the same time, to shift the value of A_μ according to

$$A_\mu \rightarrow A'_\mu = A_\mu + \frac{1}{e}\frac{\partial\theta(x)}{\partial x^\mu} \tag{22.37}$$

then eq. (22.36) will become

$$\begin{aligned}\mathcal{L}' &= \overline{\psi}(x)\left(-\gamma^\mu \frac{\partial\theta(x)}{\partial x^\mu} + i\not{\partial} + e\gamma^\mu\left(A_\mu + \frac{1}{e}\frac{\partial\theta(x)}{\partial x^\mu}\right) - m\right)\psi(x)\\ &= \overline{\psi}(x)(i\not{\partial} + e\not{A} - m)\psi(x)\\ &= \mathcal{L}\end{aligned}$$

so that the symmetry is restored. The field A_μ that restores the symmetry is called a gauge field.

In summary, imposing a local $U(1)$ symmetry on the electron field ψ requires the introduction of a gauge field A_μ in order to keep the Lagrangian invariant. The symmetry will hold if the symmetry operation (called the gauge transform) simultaneously takes

$$\psi \rightarrow \psi' \ = e^{i\theta}\psi \tag{22.38a}$$

$$A_\mu \rightarrow A'_\mu = A_\mu + \frac{1}{e}\frac{\partial\theta(x)}{\partial x^\mu} \tag{22.38b}$$

Step 3: We next show that the gauge transform (22.38) on the field A_μ is identical to the gauge transform of the scalar and vector potential functions in classical electromagnetism. In §14.8.1 it was shown that the Maxwell equations, and hence all of the physics of electromagnetism, were invariant under the gauge transform (14.21):

$$\mathbf{A} \rightarrow \mathbf{A} + \nabla\Lambda \tag{22.39a}$$

$$\varphi \rightarrow \varphi - \frac{\partial\Lambda}{\partial t} \tag{22.39b}$$

where Λ is an arbitrary function and $\mathbf{A}$ and φ are the vector and scalar potentials that give rise to the electric and magnetic fields according to eqq. (14.20a) and (14.20b):

$$\mathbf{E} = -\nabla\varphi - \frac{\partial\mathbf{A}}{\partial t} \tag{22.40a}$$

$$\mathbf{B} = \nabla \times \mathbf{A} \tag{22.40b}$$

To show the equivalence of eqq. (22.39) and eq. (22.38b), we need only identify the scalar potential φ with the time component A^0 of A^μ and the vector potential $\mathbf{A}$ with the space components A^1, A^2, and A^3 of A^μ:[50]

$$A^\mu = (A^0, A^1, A^2, A^3) \qquad \leftrightarrow \ (\varphi, \mathbf{A}) \quad = (\varphi, A_x, A_y, A_z) \tag{22.41a}$$

and consequently, according to eq. (22.26),

$$A_\mu = (A^0, -A^1, -A^2, -A^3) \ \leftrightarrow \ (\varphi, -\mathbf{A}) = (\varphi, -A_x, -A_y, -A_z) \tag{22.41b}$$

Writing out eq. (22.38b) explicitly for $\mu = 0$, we have

$$A_0 \to A_0 + \frac{1}{e}\frac{\partial\theta}{\partial x^0}$$

which, with the identifications (22.41b) for the components of A_μ, becomes

$$\varphi \to \varphi + \frac{1}{e}\frac{\partial\theta}{\partial t}$$

Likewise for $\mu = 1$ we obtain

$$\begin{aligned} A_1 &\to A_1 + \frac{1}{e}\frac{\partial\theta}{\partial x^1} \\ -A_x &\to -A_x + \frac{1}{e}\frac{\partial\theta}{\partial x} \\ A_x &\to A_x - \frac{1}{e}\frac{\partial\theta}{\partial x} \end{aligned}$$

and similarly

$$A_y \to A_y - \frac{1}{e}\frac{\partial\theta}{\partial y} \qquad\qquad A_z \to A_z - \frac{1}{e}\frac{\partial\theta}{\partial z}$$

Combining the relations for A_x, A_y, and A_z into a three-vector, we have

$$\begin{aligned} A_x\,\widehat{\mathbf{x}} + A_y\,\widehat{\mathbf{y}} + A_z\,\widehat{\mathbf{z}} &\to \left(A_x - \frac{1}{e}\frac{\partial\theta}{\partial x}\right)\widehat{\mathbf{x}} + \left(A_y - \frac{1}{e}\frac{\partial\theta}{\partial y}\right)\widehat{\mathbf{y}} + \left(A_z - \frac{1}{e}\frac{\partial\theta}{\partial z}\right)\widehat{\mathbf{z}} \\ &= (A_x\,\widehat{\mathbf{x}} + A_y\,\widehat{\mathbf{y}} + A_z\,\widehat{\mathbf{z}}) - \frac{1}{e}\left(\widehat{\mathbf{x}}\,\frac{\partial}{\partial x} + \widehat{\mathbf{y}}\,\frac{\partial}{\partial y} + \widehat{\mathbf{z}}\,\frac{\partial}{\partial z}\right)\theta \end{aligned}$$

[50] This was in fact our reason for making the contribution of A^μ to the Lagrangian $e\not{A}$ rather than just $\not{A}$; pulling out the factor of e has enabled us to identify A^μ with φ and $\mathbf{A}$, without any extraneous factors of e.

In other words, eq. (22.38b) gives

$$\varphi \to \varphi + \frac{1}{e}\frac{\partial\theta}{\partial t} \qquad\qquad \mathbf{A} \to \mathbf{A} - \frac{1}{e}\nabla\theta$$

which are indeed identical to eqq. (22.39) if we make the identification

$$\Lambda = -\frac{1}{e}\,\theta$$

Step 4: The modern sentiment is that everything mathematically possible should occur in nature. Now that we have determined that the $U(1)$ symmetry requires that a field A^μ be included in the Lagrangian, we therefore ask what other sorts of contributions this new field could make to $\mathcal{L}$; any term involving A^μ that is allowed mathematically and that does not lead to absurd or otherwise tenable physics should be included.

It turns out that there are many pathologies that can befall candidates for inclusion in the Lagrangian. Innocuous-looking terms like $A^\mu A_\mu$, for example, while Lorentz invariant, are not gauge invariant and will ruin the $U(1)$ symmetry. It can be shown that the only further contribution A^μ can make to $\mathcal{L}$ is of the form

$$-\tfrac{1}{4}F^{\mu\nu}F_{\mu\nu}$$

where the factor of $\frac{1}{4}$ is, as will be seen below, for convenience, and where $F^{\mu\nu}$, known as the electromagnetic field-strength tensor, stands for

$$F^{\mu\nu} = \frac{\partial A^\nu}{\partial x_\mu} - \frac{\partial A^\mu}{\partial x_\nu} \tag{22.42}$$

With this term, $\mathcal{L}$ becomes the full Lagrangian for quantum electrodynamics:

$$\mathcal{L} = \overline{\psi}(x)(i\partial\!\!\!/ + eA\!\!\!/ - m)\psi(x) - \tfrac{1}{4}F^{\mu\nu}F_{\mu\nu} \tag{22.43}$$

Our task is now to show that the inclusion of this $-\frac{1}{4}F^{\mu\nu}F_{\mu\nu}$ term gives rise to the Maxwell equations and hence to the electromagnetic force that we all know and love.

First, however, let us work out the components of $F^{\mu\nu}$; this will enable us to see just how it is related to the electric and magnetic fields of the Maxwell equations, and we will be needing these components later anyway. Because it has two indices, $F^{\mu\nu}$ is a 4×4 matrix and so has 16 components. Working out 16 components would be a real drag, but fortunately $F^{\mu\nu}$ is an antisymmetric matrix: as you can see from eq. (22.42), reversing the indices on $F^{\mu\nu}$ reverses its sign:

$$F^{\nu\mu} = \frac{\partial A^\mu}{\partial x_\nu} - \frac{\partial A^\nu}{\partial x_\mu} = -\left(\frac{\partial A^\nu}{\partial x_\mu} - \frac{\partial A^\mu}{\partial x_\nu}\right) = -F^{\mu\nu}$$

The four diagonal components of $F^{\mu\nu}$ (F^{00}, F^{11}, F^{22}, and F^{33}) therefore vanish, since $F^{00} = -F^{00}$ yields $F^{00} = 0$ and so on. And of the $16 - 4 = 12$ remaining components, we have only 6 to work out, since

$$\begin{aligned} F^{10} &= -F^{01} & F^{20} &= -F^{02} & F^{30} &= -F^{03} \\ F^{21} &= -F^{12} & F^{32} &= -F^{23} & F^{31} &= -F^{13} \end{aligned}$$

Using eqq. (22.41a) and (22.26) to expand the definition (22.42) of $F^{\mu\nu}$ for the case $\mu = 0$, $\nu = 1$, we have

$$F^{01} = \frac{\partial A^1}{\partial x_0} - \frac{\partial A^0}{\partial x_1} = \frac{\partial A_x}{\partial t} - \frac{\partial \varphi}{\partial(-x)} = \frac{\partial \varphi}{\partial x} + \frac{\partial A_x}{\partial t}$$

which, though with opposite sign, is just the x component of the right-hand side of eq. (22.40a). Thus

$$F^{01} = -E_x$$

and we would similarly obtain

$$F^{02} = -E_y \qquad\qquad F^{03} = -E_z$$

Again using eqq. (22.41a) and (22.26) to expand the definition (22.42) of $F^{\mu\nu}$ for the case $\mu = 2$, $\nu = 1$, we also have

$$F^{12} = \frac{\partial A^2}{\partial x_1} - \frac{\partial A^1}{\partial x_2} = \frac{\partial A_y}{\partial(-x)} - \frac{\partial A_x}{\partial(-y)} = -\left(\frac{\partial A_y}{\partial x} - \frac{\partial A_x}{\partial y}\right)$$

which, though again with opposite sign, is just the z component of the right-hand side of eq. (22.40b). Thus

$$F^{12} = -B_z$$

Exactly similar calculations would yield

$$F^{23} = -B_x \qquad\qquad F^{13} = B_y$$

Putting all this together, we have

$$\begin{aligned} F^{\mu\nu} &= \begin{pmatrix} F^{00} & F^{01} & F^{02} & F^{03} \\ F^{10} & F^{11} & F^{12} & F^{13} \\ F^{20} & F^{12} & F^{22} & F^{23} \\ F^{30} & F^{31} & F^{32} & F^{33} \end{pmatrix} \\ &= \begin{pmatrix} 0 & F^{01} & F^{02} & F^{03} \\ -F^{01} & 0 & F^{12} & F^{13} \\ -F^{02} & -F^{21} & 0 & F^{23} \\ -F^{03} & -F^{13} & -F^{23} & 0 \end{pmatrix} \end{aligned}$$

$$= \begin{pmatrix} 0 & -E_x & -E_y & -E_z \\ E_x & 0 & -B_z & B_y \\ E_y & B_z & 0 & -B_x \\ E_z & -B_y & B_x & 0 \end{pmatrix} \tag{22.44}$$

If nothing else, from the dependence of its elements on the electric and magnetic fields it should now be apparent why $F^{\mu\nu}$ is called the electromagnetic field-strength tensor.

Now back to our Lagrangian (22.43). The equation of motion (that is, the Lagrange equation) for A_ν is

$$0 = \frac{\partial}{\partial x^\mu} \frac{\partial \mathcal{L}}{\partial\left(\frac{\partial A_\nu}{\partial x^\mu}\right)} - \frac{\partial \mathcal{L}}{\partial A_\nu}$$

which, since the only term that depends on derivatives of A is the $-\frac{1}{4}F^{\mu\nu}F_{\mu\nu}$ term, and the only term that depends on just A is the $\overline{\psi}(x)e\not{A}\psi(x)$ term, reduces to

$$0 = \frac{\partial}{\partial x^\mu} \frac{\partial\left(-\frac{1}{4}F^{\mu\nu}F_{\mu\nu}\right)}{\partial\left(\frac{\partial A_\nu}{\partial x^\mu}\right)} - \frac{\partial\left(\overline{\psi}(x)e\not{A}\psi(x)\right)}{\partial A_\nu} \tag{22.45}$$

The second term on the right-hand side of eq. (22.45) works out to

$$-\frac{\partial\left(\overline{\psi}(x)e\not{A}\psi(x)\right)}{\partial A_\nu} = -\frac{\partial\left(\overline{\psi}(x)e\gamma^\mu A_\mu\psi(x)\right)}{\partial A_\nu} = -\overline{\psi}(x)e\gamma^\mu\frac{\partial A_\mu}{\partial A_\nu}\psi(x)$$

Now, the derivative will vanish unless $\mu = \nu$, in which case it is simply unity. This second term therefore reduces to

$$-\overline{\psi}(x)e\gamma^\nu\psi(x) = -e\overline{\psi}(x)\gamma^\nu\psi(x) \tag{22.46}$$

To deal with the first term on the right-hand side of eq. (22.45), we can, remembering from eq. (22.26) that the metric η can be used to raise or lower indices, rewrite [51]

$$F^{\kappa\lambda}F_{\kappa\lambda} = \eta^{\kappa\sigma}\eta^{\lambda\rho}F_{\sigma\rho}F_{\kappa\lambda}$$

Thus the first term on the right-hand side of eq. (22.45) becomes

$$\frac{\partial}{\partial x^\mu}\left(\frac{\partial\left(-\frac{1}{4}F^{\kappa\lambda}F_{\kappa\lambda}\right)}{\partial\left(\frac{\partial A_\nu}{\partial x^\mu}\right)}\right) = \frac{\partial}{\partial x^\mu}\left(\frac{\partial\left(-\frac{1}{4}\eta^{\kappa\sigma}\eta^{\lambda\rho}F_{\sigma\rho}F_{\kappa\lambda}\right)}{\partial\left(\frac{\partial A_\nu}{\partial x^\mu}\right)}\right)$$

which, if we use eq. (22.42) and note that η, since it is a constant matrix, can be pulled outside of the derivatives, works out to

$$= -\frac{1}{4}\eta^{\kappa\sigma}\eta^{\lambda\rho}\frac{\partial}{\partial x^\mu}\left(\frac{\partial(F_{\sigma\rho}F_{\kappa\lambda})}{\partial\left(\frac{\partial A_\nu}{\partial x^\mu}\right)}\right)$$

[51] We use indices κ and λ (and σ and ρ) here rather than μ and ν because we will be plugging this into eq. (22.45), which already uses μ and ν.

$$= -\tfrac{1}{4}\eta^{\kappa\sigma}\eta^{\lambda\rho}\frac{\partial}{\partial x^\mu}\left(\frac{\partial\left(\left(\frac{\partial A_\sigma}{\partial x^\rho}-\frac{\partial A_\rho}{\partial x^\sigma}\right)\left(\frac{\partial A_\kappa}{\partial x^\lambda}-\frac{\partial A_\lambda}{\partial x^\kappa}\right)\right)}{\partial\left(\frac{\partial A_\nu}{\partial x^\mu}\right)}\right)$$

We will therefore have four contributions to the derivative with respect to $\partial A_\nu/\partial x^\mu$: one from the case $\sigma=\nu$ and $\rho=\mu$, one from the case $\rho=\nu$ and $\sigma=\mu$, one from the case $\kappa=\nu$ and $\lambda=\mu$, and one from the case $\lambda=\nu$ and $\kappa=\mu$. For the case $\sigma=\nu$ and $\rho=\mu$, the derivative hits the first of the four derivatives of A in the numerator and gives a contribution of

$$\begin{aligned}-\tfrac{1}{4}\eta^{\kappa\nu}\eta^{\lambda\mu}\frac{\partial}{\partial x^\mu}\left((+1)\left(\frac{\partial A_\kappa}{\partial x^\lambda}-\frac{\partial A_\lambda}{\partial x^\kappa}\right)\right) &= -\tfrac{1}{4}\eta^{\kappa\nu}\eta^{\lambda\mu}\frac{\partial}{\partial x^\mu}(F_{\lambda\kappa})\\ &= -\tfrac{1}{4}\frac{\partial}{\partial x^\mu}(\eta^{\kappa\nu}\eta^{\lambda\mu}F_{\lambda\kappa})\\ &= -\tfrac{1}{4}\frac{\partial}{\partial x^\mu}(F^{\mu\nu})\end{aligned}$$

The other three cases will give exactly the same contribution. For the case $\rho=\nu$ and $\sigma=\mu$, for example, the derivative will hit the second of the four derivatives of A in the numerator, which would appear to give a contribution differing by a sign, but outside we will now have $\eta^{\kappa\mu}\eta^{\lambda\nu}$ instead of $\eta^{\kappa\nu}\eta^{\lambda\mu}$ and will therefore end up with an $F^{\nu\mu}$ rather than an $F^{\mu\nu}$. Since $F^{\nu\mu}=-F^{\mu\nu}$, this restores the sign. The remaining two cases work out similarly. So for the total contribution of the first term on the right-hand side of eq. (22.45) we need simply multiply the above contribution by 4:[52]

$$-\frac{\partial F^{\mu\nu}}{\partial x^\mu} \tag{22.47}$$

Using our results (22.47) and (22.46) for the first and second terms on the right-hand side of the equation of motion (22.45) for the field A^μ we arrive at

$$0 = -\frac{\partial F^{\mu\nu}}{\partial x^\mu} - e\overline{\psi}(x)\gamma^\nu\psi(x)$$

or

$$\frac{\partial F^{\mu\nu}}{\partial x^\mu} = -e\overline{\psi}(x)\gamma^\nu\psi(x) \tag{22.48}$$

While we cannot work through all the details here, it should seem plausible that $-e\overline{\psi}(x)\gamma^\nu\psi(x)$ is the four-vector current density j^ν,

$$j^\nu = (\rho, \mathbf{j})$$

where ρ and $\mathbf{j}=\rho\mathbf{v}$ are the charge and current densities that occur in the Maxwell eqq. (22.24): the combination $\overline{\psi}\psi$ is similar to the absolute square

[52] This factor of 4 is in fact the reason for the $\frac{1}{4}$ in $-\frac{1}{4}F^{\mu\nu}F_{\mu\nu}$.

$|\psi|^2$ in quantum mechanics, which corresponded to the probability density (the probability per unit volume) of finding the electron at spacetime location x. Multiplying the electron's probability per unit volume by its charge $-e$ should then give the corresponding charge ρ per unit volume. Although it is far from obvious and well beyond what we can get into, it turns out that the γ^ν gives a factor of the velocity $\mathbf{v}$ for the spatial components of $-e\overline{\psi}(x)\gamma^\nu\psi(x)$, which therefore give not just ρ, but $\rho\mathbf{v} = \mathbf{j}$. Eq. (22.48) thus becomes

$$\frac{\partial F^{\mu\nu}}{\partial x^\mu} = j^\nu \tag{22.49}$$

with

$$j^\nu = (j^0, j^1, j^2, j^3) \quad \leftrightarrow \quad (\rho, \mathbf{j}) = (\rho, j_x, j_y, j_z) \tag{22.50}$$

It remains only to re-express eq. (22.49) in terms of the electric and magnetic fields, which we can easily do using eq. (22.44). For the case $\nu = 0$, eq. (22.49) becomes

$$\begin{aligned}
\frac{\partial F^{\mu 0}}{\partial x^\mu} &= j^0 \\
\frac{\partial F^{00}}{\partial x^0} + \frac{\partial F^{10}}{\partial x^1} + \frac{\partial F^{20}}{\partial x^2} + \frac{\partial F^{30}}{\partial x^3} &= j^0 \\
0 + \frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} &= \rho \\
\nabla \cdot \mathbf{E} &= \rho
\end{aligned}$$

which is just Gauss's law (22.24a). Similarly for the case $\nu = 1$, eq. (22.49) becomes

$$\begin{aligned}
\frac{\partial F^{\mu 1}}{\partial x^\mu} &= j^1 \\
\frac{\partial F^{01}}{\partial x^0} + \frac{\partial F^{11}}{\partial x^1} + \frac{\partial F^{21}}{\partial x^2} + \frac{\partial F^{31}}{\partial x^3} &= j^1 \\
\frac{\partial(-E_x)}{\partial t} + 0 + \frac{\partial B_z}{\partial y} + \frac{\partial(-B_y)}{\partial z} &= j_x \\
-\frac{\partial E_x}{\partial t} + \frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} &= j_x
\end{aligned}$$

This is just the x component of

$$-\frac{\partial \mathbf{E}}{\partial t} + \nabla \times \mathbf{B} = \mathbf{j}$$

which is equivalent to eq. (22.24d). The cases $\nu = 2$ and $\nu = 3$ similarly yield the y and z components of this relation.

The equation of motion for the field A^μ is thus equivalent to two of the four Maxwell equations. What about the other two? As we saw in §14.8, when the electric and magnetic fields are written in terms of the potential functions φ and $\mathbf{A}$, these other two equations, (22.24b) and (22.24c), become identities that are automatically satisfied:

$$\nabla \cdot \mathbf{B} = \nabla \cdot (\nabla \times \mathbf{A}) = 0$$
$$\nabla \times \mathbf{E} = \nabla \times \left(-\nabla\varphi - \frac{\partial \mathbf{A}}{\partial t} \right) = 0 - \frac{\partial}{\partial t}(\nabla \times \mathbf{A}) = -\frac{\partial \mathbf{B}}{\partial t}$$

In summary, we start with the physics — that is, with the Lagrangian — of a free electron field ψ. If we impose on this field a local $U(1)$ symmetry, equivalent to supposing that associated with every point in spacetime there is a little circle and that one's orientation or position on this little circle has no effect on the electron field, then a new field A^μ must be introduced. The most general contribution this new field can make to the Lagrangian, $-\frac{1}{4}F^{\mu\nu}F_{\mu\nu}$, then exactly reproduces the Maxwell equations and all of electromagnetism. The $U(1)$ symmetry of the electron field thus gives rise to the electromagnetic force.

If you were a god trying to construct a universe out in your garage, your reasoning would run something like this: a universe with particles only of integer spin wouldn't be very interesting, because there is no limit to the number of integer-spin particles that can occupy a given state — matter would just settle down into a boring soup near the ground state. To make interesting stuff, you need matter that can be stacked up like Legos, and this would lead you to create particles of half-odd integral spin (that is, spin $\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, ...), only one of which can occupy a given state.[53] The simplest particle of half-odd integral spin you could come up with would be something like the spin-$\frac{1}{2}$ electron. But just an electron field by itself still wouldn't be very interesting: you can stack the electrons up, but they don't interact with each other, so nothing interesting would happen. The simplest way to spice things up would be to impose a symmetry like that of the unit circle on the electron field — if you were God, it's the very first thing you would think to do. And once you impose this symmetry, presto, you have the Maxwell equations and the familiar electromagnetic force. Absolute child's play.

Anyway, in quantum theory, the field A^μ is the photon. The $\overline{\psi}(x)(i\partial\!\!\!/ - m)\psi(x)$ part of the Lagrangian corresponds to the electron-positron propagator — the mathematical machinery governing the motion of free electrons and positrons. In Feynman diagrams, this propagator corresponds to the straight lines that represent electrons and positrons (as in, for example, fig. (22.1) on

[53] From chemistry, you may be used to thinking of two electrons per orbital, but that's because you were not counting the spin-up and spin-down states separately.

p.1074). The $-\frac{1}{4}F^{\mu\nu}F_{\mu\nu}$ term in the Lagrangian corresponds to the photon propagator and is represented in Feynman diagrams by wavy lines. The remaining term, the $e\overline{\psi}(x)\not{A}\psi(x)$, corresponds to the interaction between electrons (or positrons) and photons. In Feynman diagrams, this interaction is the vertex at which the electron-positron and photon lines meet. Because this term involves two electron-positron fields ($\overline{\psi}$ and ψ) and one photon field, such vertices always have two electron lines and one photon line. This simple vertex is the only way that electrons and photons can interact with each other: an electron can emit a photon or absorb a photon, and that's it. But that's enough to give rise to all of the myriad manifestations of the electromagnetic interaction that we observe in nature.

The elements of the group $U(1)$ were of the form $e^{i\theta}$ and thus have the property that their inverses are their complex conjugates:

$$e^{i\theta}(e^{i\theta})^* = e^{i\theta}e^{-i\theta} = 1$$

Closely related to $U(1)$ are the groups $SU(n)$, which consist of all $n \times n$ matrices of unit determinant whose inverses are their Hermitian conjugates:

$$u \in SU(n) \quad \Longrightarrow \quad \det u = 1, \quad uu^\dagger = 1$$

For a matrix, Hermitian conjugation is the equivalent of complex conjugation for ordinary numbers: to take the Hermitian conjugate of a matrix, you take its complex conjugate and then its transpose. The S and U in SU stand for *special* and *unitary*, respectively, with *special* meaning *having unit determinant* and *unitary* meaning that the matrix's inverse is its Hermitian conjugate (the $uu^\dagger = 1$ property). The group $SU(2)$ turns out to correspond to the weak interaction involving quarks and leptons (electrons, positrons, muons, taus, and their associated neutrinos). Since all of the elements of $U(1)$ can be expressed in the single form $e^{i\theta}$, there is only one corresponding gauge field A, which corresponds to the photon. To express the elements of $SU(2)$, three matrices are required, and the three gauge fields corresponding to them turn out to be the W^+, the W^-, and the Z particles that convey the weak interaction. The symmetry of the strong interaction between quarks and gluons turns out to be $SU(3)$. To express the elements of $SU(3)$, eight matrices are required, and the eight gauge fields corresponding to them are the gluons. The symmetry of gravity is local Poincaré invariance, that is, invariance under local Lorentz transforms and translations in four-dimensional spacetime. So we have a simple hierarchy of symmetries: $1 + 2 + 3 + 4 = 10$. It turns out that bowling was more profound than you thought.

22.7 Unification Theories

There have been various attempts to unify the four interactions and find the single, underlying symmetry of the universe. Here we outline some of the conceptually more important ones.

22.7.1 Kaluza-Klein Theories

As noted in §22.6 and shown explicitly in §22.6.1, the $U(1)$ symmetry that gives rise to the electromagnetic interaction can be represented by the set of complex phases $e^{i\varphi}$, or, equivalently, a circle. The symmetry that gives rise to the gravitational interaction in general relativity is local Poincaré invariance in four-dimensional spacetime. Mathematically, one works with a 4×4 matrix formalism — the metric of §10.9 and the associated mathematical machinery leading to the gravitational field equations.

In 1921 Theodor Kaluza took Einstein's general theory of relativity and tried something really outlandish: just to see what would happen, he extended spacetime to five dimensions by adding a fourth spatial dimension. Mathematically, this meant extending the metric from a 4×4 to a 5×5 matrix. When he then worked out the Einstein field equations in this five-dimensional spacetime, lo and behold he obtained not just the usual gravitational field equations of four-dimensional spacetime, but also, as a direct consequence of the extra dimension, the Maxwell equations of electromagnetism.

In somewhat more detail, what happens is this: extending a 4×4 matrix to 5×5 means adding an extra diagonal element, plus the four new elements in the column above it, plus the four new elements in the row in front of it. That would be nine extra parameters. But the metric is a symmetric matrix, so the four new "column" elements have to be the same as the four new "row" elements and the total number of new parameters is actually only five. Of these, the four new "row-column" elements turn out in essence to correspond to the four-vector potential $A^\mu = (\varphi, \mathbf{A})$ of electromagnetism.[54]

As nifty and powerful as Kaluza's theory was, it immediately raised the question, Why aren't we aware of this fourth spatial dimension? Why can't we see and move around in it like we can the three familiar spatial dimensions? To address this and certain other technical problems with Kaluza's theory, in 1926 Oskar Klein suggested that the extra fifth spatial dimension was, unlike the three spatial dimensions that we do experience, curled up in a tiny loop about 10^{-33} cm in size. That is, it is as though there is a little loop associated with every point in our familiar four-dimensional spacetime. While it is not possible to visualize extending the three spatial dimensions

[54]The fifth new parameter corresponds to a scalar field known as the dilaton that is more or less of a sideshow and that we will therefore not discuss.

with which we are familiar,[55] you can visualize extending a two-dimensional plane by adding a curled-up extra dimension: it would be like a carpet, with each point (x, y) in the plane having a loop of thread attached to it. If you were as tiny as the loop, you would experience all three spatial dimensions — you could walk around the loop as well as in the x and y directions. But when you are far larger than the loops, they are too small to be directly visible to you; from your perspective there is only a two-dimensional plane with x and y directions. And so it is with the loops of Kaluza-Klein theory: at 10^{-33} cm, they are twenty orders of magnitude smaller than a nucleus, hopelessly beyond the realm of direct observation. But this loopy extra dimension corresponds exactly to the circular $U(1)$ symmetry needed to reproduce the Maxwell equations and the familiar electromagnetic force.

It is possible, by adding still more loopy extra dimensions, to construct Kaluza-Klein theories that also give rise to the weak and strong interactions. Sort of. Unfortunately, it turns out to be very difficult for Kaluza-Klein theories accommodate one essential feature of the physics of our universe: we know that nature violates parity, that is, that fermions are chiral, that is that is, that the weak interaction distinguishes between particles of right- and left-handed spin. But though Kaluza-Klein theories have largely fallen by the wayside as candidates for the ultimate physical theory, the idea of extra dimensions giving rise to forces is too beautiful not to have some validity, and this mechanism in fact turns out to be natural to more promising theories like string theory.

22.7.2 Grand Unified Theories

Known by the acronym GUTs, these theories attempt to extend the Standard Model by finding a still higher symmetry that unifies the strong with the electroweak interaction at high energies and that spontaneously breaks at lower energies. Many schemes for doing this have been proposed, but all are problematic in one respect or another. In particular, the proton, which is stable (that is, never decays) in the Standard Model, is predicted by most GUTs to decay at rates already ruled out experimentally: very sensitive and long-running experiments have not only established a lower limit of some 10^{35} yr for the lifetime of the proton, but have failed to detect even a single proton decay. Some other GUTs predict relationships among particle masses that are inconsistent with observation. While not all GUTs have been ruled

[55] There are some charlatans who claim to be able to visualize four or more spatial dimensions, but what they're really doing is thinking of projections of those higher dimensions onto three spatial dimensions, which is just a cheap trick. You can't any more visualize extra dimensions than you can imagine colors beyond those in the rainbow. But that of course is no impediment to reasoning about with extra dimensions logically and dealing with them mathematically.

out, even if otherwise successful these theories still would not include gravity and therefore cannot be considered candidates for the ultimate physical theory.

22.7.3 Supersymmetry & Supergravity

The search for a higher symmetry that would include all of the symmetries in nature was confronted with a major impediment in 1967, when Sidney Coleman and Jeffrey Mandula were able prove mathematically what has come to be known, naturally enough, as the Coleman-Mandula theorem: there are no nontrivial ways to combine the Poincaré symmetries of Lorentz transforms and spacetime translations with internal gauge symmetries of the kind found in the Standard Model. The proof of the theorem relies, however, on an assumption that the parameters of the symmetries are ordinary complex numbers; in the early 1970s, it was discovered that if the parameters are instead taken to include Grassman variables — anticommuting variables for which $ab = -ba$ — it turns out that another symmetry is possible. This new, higher symmetry is called supersymmetry, or SUSY for short.

Mathematically, the spins of particles must be integral (0, 1, 2, ...) or half-odd integral ($\frac{1}{2}$, $\frac{3}{2}$, $\frac{5}{2}$, ...). Particles of integral spin, such as the photon and the gluon, are known as bosons, particles of half-odd integral spin, such as the electron and the quarks, as fermions, and these two types of particles behave very differently. For example, while there is no limit on the number of bosons that can occupy a given quantum state, each state can hold only one fermion. The quantum field theory of the Standard Model has no explanation for the difference in the properties of bosons and fermions, nor is there any symmetry linking the two.

Supersymmetry, however, encompasses both fermions and bosons — its symmetry operation transforms them into each other in the same way that the Lorentz transform mixes spatial displacements and time intervals. And if that symmetry, instead of being global (that is, applying uniformly to all points in spacetime), is made local (so that the symmetry applies independently to each point in spacetime), supersymmetry turns out to incorporate gravity. Theories having local supersymmetry are thus called supergravity (or SUGRA, for short). Supergravity would not only unify all four interactions, but is free of most of the nasty infinities that plague nonsupersymmetric quantum field theory, thus eliminating the need for many of the contortions of renormalization. More importantly, supergravity would, if we had the math to work it out, also provide a viable quantum theory of gravity.

Among the problems with SUSY and SUGRA is that they predict not only the observed particles, but hordes of other, unobserved particles: for every currently observed fermion, there is an as yet unobserved bosonic partner called a sfermion, and for every currently observed boson, there is an as

yet unobserved bosonic partner called a bosino: there are "gluinos," "Z-inos," and "W-inos" corresponding to gluons, Z's, and W's; "selectrons" and "squarks" corresponding to electrons and quarks; etc.[56] While it is possible that these unobserved particles are all simply too massive to have been observed to date, there is not yet a compelling explanation for such a lopsided asymmetry in the masses of the hitherto observed particles and their as yet unobserved supersymmetric partners. But that certainly doesn't rule the theory out.

And that is good, because SUSY and SUGRA are integral to the currently most promising theory, string theory. In fact we may even be on the verge of experimental confirmation of the existence of supersymmetric particles: for years a team at Brookhaven has been making ever finer measurements of the gyromagnetic ratio of the muon, and recently they have found contributions to it that are of the size and sort that would be expected theoretically from virtual supersymmetric particles. At present, there has been no definite conclusion; the theoretical calculations, which have to be carried out to many orders in perturbation theory, are extremely complex, and no one can be confident at this point that they are free of mistakes.

22.7.4 String Theory

Some believe that string theory may prove the ultimate physical theory; that it will not only completely but very possibly uniquely account for the physics of the universe.[57]

String theory is similar to quantum field theory; the difference is the dimensionality of the fundamental objects. In quantum field theory, the basic physical entity, the particle, is point-like: the value of the quantum field at a point in spacetime tells us how many particles of that type are at that point. In string theory, the basic physical entity is one-dimensional, a string — specifically, a closed string in the form of a little loop about 10^{-33} cm in size. Such small loops are essentially point-like, so string theory will reproduce quantum field theory and the Standard Model at low energies, but it turns out that string theory incorporates all the best features of other unification theories while being free from their pathologies. Moreover, the theory is remarkably simple: the action (in the sense, as discussed in Chapter 21, of the integral of the Lagrangian) for the heterotic string — the single equation

[56] As you may have gathered, for the bosonic partner of a fermion you prefix an *s-* to the name of the fermion, and for the fermionic partner of a boson you add an *-ino* suffix to the name of the boson.

[57] And even if string theory doesn't pan out, one of the really nice things about being a theoretical physicist is that you can destroy the entire universe and then, with a little bit of eraser, everything is okay again.

to which all of physics might reduce — can be written as

$$S = -\frac{1}{2\pi}\int d^2\sigma\,\sqrt{-h}\Big(h^{\alpha\beta}\partial_\alpha X^\mu \partial_\beta X_\mu + i\overline{\psi}^\mu \rho^\alpha \partial_\alpha \psi_\mu\Big)$$

where the X^μ are the ten-dimensional coordinates of the string, the ψ^μ are the fermionic fields, and the integration by $d^2\sigma$ is over a two-dimensional (1 space + 1 time) background spacetime in which the strings "live." Note the simplicity of this string action compared to the sprawling mess of the Standard Model Lagrangian (22.23) of p.1077 — which, unlike the string action, does not even include gravity.

Among the very nice features of string theory:

- String theory doesn't suffer from the nasty infinities of quantum field theory and doesn't require renormalization.[58]
- String theory includes all four interactions — the gravitational as well as the electromagnetic, weak, and strong.
- String theory is capable of being broken down at low energies into the various symmetries of the Standard Model, as well as giving rise to the three generations of quarks and leptons observed in the universe.
- String theories work in ten dimensions — exactly what we need for the physics we observe in the universe: $1+2+3+4=10$ for the electromagnetic $U(1)$, the weak $SU(2)$, the strong $SU(3)$, and the Poincaré invariance of four-dimensional spacetime.

  Much more compellingly, quantum string theory can be formulated in a logically consistent way only in exactly ten dimensions.[59] Thus it isn't

[58] This finiteness is a consequence of the strings being one-dimensional rather than point-like. The nasty infinities in quantum field theory at high energies and momenta correspond to short distances, as you can see from our complex exponential for a wave function, $e^{i(px-Et)/\hbar}$: when p is very large, the exponential oscillates very rapidly, corresponding to a very short wavelength and therefore to a very short distance scale. The divergences as $p\to\infty$ correspond to $x\to 0$, that is, to shrinking down to a point. Since the fundamental objects in quantum field theory are point-like, bad things happen in this limit, similar to the divergence of the electric field

$$\frac{1}{4\pi\epsilon_0}\frac{q}{r^2}$$

of a point charge as $r\to 0$. Strings, being spread out over small loops rather than point-like, do not share this pathology.

[59] More specifically, it turns out that there is no sensible way to define the operators that generate Lorentz transforms unless the number of dimensions equals ten. Believe it or not, the proof of this makes use of the sum of the positive integers:

$$\sum_{n=1}^{\infty} n = 1+2+3+4+\cdots = -\tfrac{1}{12}$$

as is demonstrated, for those skeptical, in Appendix E.

simply that string theories are possible in the ten dimensions needed to reproduce the four interactions we observe, it's that string theories are *only* possible in this needed number of dimensions.

- As opposed to the 19 or so arbitrary parameters of the Standard Model, in string theory there is only one arbitrary parameter, known as the string tension; all other physical parameters — the strengths of the four interactions, the various particle masses, etc., are determined by the theory, including dimensionless parameters like the fine structure constant $\alpha = e^2/4\pi\epsilon_0\hbar c$.[60]

- For a time it seemed that there were five distinct string theories, which would leave us with the question of how our universe chose among them. It was then discovered that there are pairwise relations (known as dualities) among these five theories and that they are in fact just different representations of a single underlying eleven-dimensional theory called M-theory.[61] It therefore appears that string theory is unique.

In summary, it looks like string theory may not only fully and uniquely account for the physical universe, but may be the *only possible theory for a physical universe* — any other theory would be internally logically inconsistent. In other words, string theory may explain not only why the physical universe is the way it is, but why the physical universe could not have been otherwise.[62]

At present this is, however, all rather speculative; string theory is so different from previous theories that not only new techniques but also new mathematics is needed to work through it — string theorists find themselves having to invent the required math as they go along, much the way Newton had to invent calculus to formulate his theories of dynamics and gravitation. So although things look very promising, the theory is far from having been fully worked out. And while the theory has been shown to be unique, unfortunately its vacuum state is far from being so: there are estimated to be

[60] The agreement of such predictions with the known empirical values of these various quantities would, of course, confirm the correctness of string theory to even the most adamant skeptic's satisfaction. Unfortunately, while in principle we know that all these quantities are calculable in string theory, we are still far from being able to carry out such calculations. But stay tuned.

[61] One might make a very rough analogy to straight lines, circles, ellipses, parabolas, and hyperbolas: these seem like five quite different sorts of curves in a two-dimensional plane, but when you embed them in three dimensions, you see that they are all just different slicings through a cone.

[62] Or this may also turn out to be a lie. It may be that string theory is just another in a series of increasingly better approximations to the real truth, to the ultimate physical theory. Or maybe there isn't any truth after all; maybe physics is just a series of vicious lies and physicists have just been having a lot of fun at the expense of everyone else. Or maybe God is the joker and the universe really doesn't make any sense at all. Search us.

some 10^{500} possible vacuum states, each giving rise to a physics more or less different from those of the others. This embarrassment of riches leads some to argue that string theory lacks predictive power and therefore to question whether it even constitutes a scientific theory. But there is no reason to believe that we will not, in time, discover some principle that will make the choice unique or nearly so. And even if no such principle is discovered, the cosmological mechanism of eternal inflation[63] — with its bubble universes so unimaginably numerous that all possibilities, no matter how remote, must be realized — could still explain how we just happen to find ourselves in our particular vacuum state.

In our opinion — and everyone knows how much our opinion is worth[64] — string theory is far too beautiful and works far too well not to be a huge step toward the ultimate physical theory. It cannot be coincidence that the theory works only in exactly the number of dimensions required to reproduce the physics we observe in the universe, and its uniqueness argues equally strongly in its favor. But string theory cannot be the final word because it still leaves some aspects of nature unexplained. In spite of the depth and breadth to which we have penetrated the workings of the physical universe, we still have, for example, no understanding of the nature and unidirectionality of time. String theory, while it constrains the number of spacetime dimensions, takes spacetime as a given, whereas spacetime and its properties should emerge from a truly fundamental theory rather than be assumed by it. String theory also assumes that the universe obeys the principle of least action, and, to borrow a phrase from I.I. Rabi, who ordered that?[65]

22.7.5 The Empirical Myth?

Traditionally physics has been regarded as an experimental science, meaning that a theory can only be formulated from experimental evidence, and, once formulated, can only be verified by experimental evidence. In this traditional view, physics is regarded as a recurring sequence of experimental observation, theoretical hypothesis, and experimental verification, with experimentation both beginning and ending the process. There are those who would maintain that our theoretical understanding of physics is only as good as the latest experiment, who regard theory as little more than an asymptotic series of

[63] Eternal inflation is explained in §22.8.

[64] We were once offered a penny for our thoughts. And that was years ago; what with inflation, that would be well over 2¢ today.

[65] Rabi actually asked this question of the discovery of the muon: at the time, quantum theory had just reached the point of nicely explaining atoms and the physics of the then-familiar particles like the electron; the unexpected discovery of the muon suddenly presented a whole new mystery.

progressively better approximations to the ultimate truth — if there even is an ultimate truth.

Historically this view evolved in a very natural way. Initially, our understanding of the physical world was, like that of an infant, limited to phenomena we directly experienced — for example, you push something and it moves. Our nascent understanding of the resulting motion being as yet too rudimentary to suggest anything further about its nature, our only means of advancing our understanding was through additional experimental investigation. Even when our understanding eventually became strong enough to suggest other properties the motion might have, we could not be sure our conjectures were correct without testing them experimentally. If experiment verified our predictions, we took our understanding to be correct; if not, we revised our understanding accordingly. And so we have forged ahead, right up to the present day, blundering through the dark, relying on experiment to suggest directions to take when we are at a loss or to confirm conjectures of which we aren't entirely sure — with the result that physics has come to be viewed as an experimental science.

Historical evolution and rational justification are, however, two very different things, and the view that physics is properly an experimental science, which has never had any rational justification, may well prove merely an unscientific prejudice. String theory now presents us with the possibility that the physics of our universe may in fact be uniquely determined, that is, that there is only one logically self-consistent physical theory and that therefore the physical universe is the way it is because there is no other way it could be. If this does prove to be the case, then experimentation will have been only a temporary crutch that we needed to limp along with our as yet incomplete understanding and, quite contrary to the traditional view, physics will turn out to be properly an entirely theoretical venture.[66]

Either way, there remains the great mystery of why the universe is such a profoundly mathematical place. As P.A.M. Dirac observed, "It seems that if one is working from the point of view of getting beauty in one's equations, ... one is on a sure line of progress." Indeed, with the possible exception of pure number theory, there is no branch of mathematics that has not turned out to have physical applications, to the point that modern physics takes the perspective that anything mathematically possible should exist or occur physically. But why should this be so? One can easily understand the Darwinian origins of basic arithmetic logic: if I have three apples and club Zog, who has four apples, I will have seven apples to feed the family perpetuating

[66]Though there may remain some physical parameters the values of which are determined by quantum randomness and therefore have to be measured. And all of this is, of course, based on the assumption that the universe has to be logically self-consistent. But if this assumption is wrong, there is no point to trying to make sense of the universe at all.

my genes. And yet even when pursued far beyond what has any business applying to the world in which we live — to imaginary numbers involving $i = \sqrt{-1}$, to Grassmann variables for which $ab = -ba$, to the sum of the positive integers equalling $-\frac{1}{12}$, and to all sorts of other counterintuitive abstractions — , we find that mathematics still governs nature.[67] It begins to seem not so much that mathematics is true to nature as that nature is true to mathematics.

22.8 Our Amazing & Expanding Universe

To see what the early universe, and perhaps even its origin, were like, you take the state of the universe as presently observed and use physical theories like the quantum field theory and general relativity to extrapolate backward in time. The picture we put together from this backward extrapolation is something like this:[68]

In the beginning, there was *nothing* — not only no matter, but also no space and no time. Spacetime is actually a part of the universe, not something the universe is embedded in. The question, What was there before the universe? is therefore a stupid question — there wasn't any time. Similarly stupid is the question, What's outside the universe? So if you want to look sharp, don't ask those kinds of questions.[69]

The universe began as a vacuum bubble of the sort discussed on p.1076. In quantum field theory, vacuum bubbles occur all the time. The difference here would be that spacetime would be part of what comes into existence with the bubble — it would just be four of the bubble's ten dimensions.[70]

[67]This is very different from the situation with any number of other intellectual abstractions, such as literary criticism, economics, and especially philosophy, which always starts from a soundly good idea and, by extrapolating from it, invariably arrives at a philosophical system that is obviously — to everyone except philosophers — laughably absurd.

[68]In the interest of space and time (*haw! haw!*), the skeletal outline presented here is an especially gross oversimplification and, because of the considerable license inevitably required for such oversimplification, necessarily somewhat misleading and inaccurate. Many major issues are not even mentioned. But, hey, what do you expect in three-odd pages? If you want depth and thoroughness, check out the books by Greene, Guth, and Weinberg listed at the beginning of this chapter. Or if you want a nice, concise (about 15 pages) overview of cosmology that isn't really mathematical but is beyond a popular-science article, see Verde, *A Taste of Cosmology*, `https://arxiv.org/abs/1504.05945`.

That said, as hokey as some of it may seem, standard cosmology is actually pretty soundly established both theoretically and experimentally. Different sources may give somewhat differing details, but there is general agreement on the overall picture.

[69]If for some reason you are not satisfied with this advice, you can find a very thorough, interesting, and lucid discussion of space and time in Greene's *Fabric of the Cosmos*.

[70]We are assuming here that ultimately we will have a theory from which spacetime

At this early stage, all ten of the universe's dimensions were of comparable size, and the universe exhibited its full symmetry. Your grandparents may have told you stories of that time.

Anyway, at this point the vacuum-bubble universe was living on energy borrowed from itself — pulling itself up by its own bootstraps, as it were. The overwhelming odds were that the universe would simply swallow itself back up and return to nothingness. Obviously that didn't happen, so we should probably feel pretty fortunate.

Without a quantum theory of gravity, our understanding of the first 10^{-43} sec is a bit sketchy, but by that time spontaneous symmetry breaking would have separated gravity from the other interactions, and solutions to the equations of general relativity indicate that the universe would have tended to expand. That is, the three familiar spatial dimensions would have tended to get larger, while the six other spatial dimensions (the ones that give rise to the electromagnetic, weak, and strong interactions) remained small. As the universe expanded, it cooled as its energy density lowered, and by about 10^{-36} sec it had cooled enough that the strong interaction separated from the electroweak interaction by spontaneous symmetry breaking.[71] From more or less about 10^{-35} to 10^{-33} sec, during a period called inflation, the size of the universe began to grow exponentially, at much faster than the speed of light.[72] This might seem like a violation of relativity, but although no object can travel through spacetime faster than the speed of light, there is no reason why spacetime itself cannot expand faster than the speed of light. The classic analogy is to a balloon: the expansion of the universe is like that of a balloon skin as the balloon is inflated. Although no objects on the balloon can travel over its surface at a speed faster than that of light, if the balloon is inflated rapidly enough the distance between some points on its surface may grow faster than the speed of light.[73] During the brief 10^{-33} sec that inflation lasted the expansion of the universe was so rapid that it increased in size by a whopping factor of some 10^{30}, to about the size of a

emerges naturally rather than being taken as a given. Since we do not at present have such a theory — even in string theory the existence of spacetime is simply assumed — this is actually going beyond standard cosmology.

[71] For an explanation of spontaneous symmetry breaking, see p.1071.

[72] For much more about inflation, see Guth's *The Inflationary Universe* and §22.8.1, where we work through the relations that give rise to inflation.

[73] It is important to understand that this analogy between spacetime and the balloon applies only to the two-dimensional surface of the balloon: while the balloon's skin is expanding radially outward into the surrounding, pre-existing three-dimensional space, the universe is not expanding, as though its matter were exploding outward, into any pre-existing space; it is space itself that is expanding.

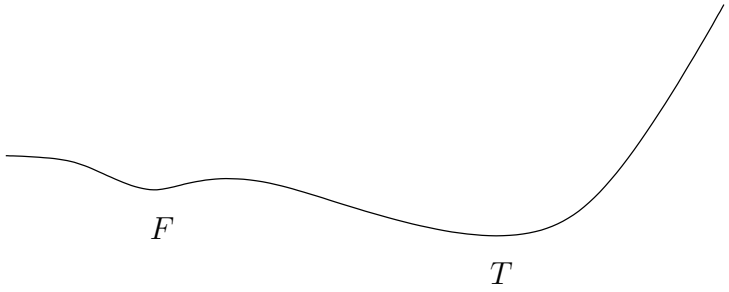


Figure 22.3: A Potential with a False Vacuum

grapefruit. Or possibly a pomegranate. Or maybe a pomegranate seed.[74, 75]

Anyway, the particle responsible for the inflation of the universe is called the inflaton. There remains much that is not known about the inflaton, but we do know that as the universe was undergoing its inflationary expansion and cooling down, it found itself in a supercooled state. That is, the field configuration of the inflaton was such that the universe found itself sitting at a point on the inflaton's potential energy curve that was an equilibrium point, but not the point of lowest energy. Such a situation is illustrated in fig. (22.3): the universe found itself at the equilibrium point F, which was at a higher energy than the equilibrium point T. Points like F are known as false vacua; the true vacuum is the state of lowest energy, where the potential is at its global minimum. The false vacuum in which the universe found itself may, as shown in fig. (22.3), have been in a small valley, separated from the true vacuum by a slight ridge, or it may have been atop a broad peak; in either case, the subsequent descent to the true vacuum T was, as shown in fig. (22.3), very gradual. Although in classical physics an object at rest at an equilibrium point will remain there forever, at some point during inflation the universe was set on a slow "roll" from the false to the true vacuum by a random quantum fluctuation (tunneling, if necessary, through any ridge that may have separated the vacua). The consequent release of energy reheated the universe, ending inflation and giving rise to all of the matter we see around us.[76] Before the transition to the true vacuum, the universe was

[74] Or even smaller. Or larger. Estimates of the size of the universe by the end of the inflationary period depend very sensitively on imprecisely known parameters and therefore vary considerably. But this is no blemish on inflationary theory; though it might seem otherwise, precisely how certain numbers like this work out is a marginal detail.

[75] A quantum fluctuation — a random subatomic fluctuation or variation due to the probabilistic nature of quantum theory — is one of the smallest-scale phenomena in the universe. Such fluctuations occur around us all the time, undetected because of their microscopic extent. As a curious result of inflation, the quantum fluctuations in the energy density of the early universe now appear in the form of clusters of galaxies — the largest-scale structures in the universe.

[76] Actually, inflation may not have ended: that the entire universe made the transition

living on borrowed energy; after the transition, we found ourselves in an expanding but stable universe filled with matter from which life could later develop. It was like we were getting our money for nuthin' and our chicks for free. Alan Guth, who first formulated inflationary theory, has characterized the universe as "the ultimate free lunch."[77]

As the universe continued to expand after the end of the inflationary period, it had cooled enough by about 10^{-12} sec for the electromagnetic and weak interactions to separate from each other by spontaneous symmetry breaking. As the universe expanded and cooled still further, quarks condensed into nuclei, then electrons fell into orbits around these nuclei and formed atoms. What had previously been an electromagnetically opaque soup of loose nuclei, electrons, and photons suddenly settled into a relatively transparent universe of atoms. The photon radiation left over from that still soup permeates the universe and, after further cooling as the universe continued to expand, has become the ubiquitous 3°K microwave background that we see today.[78]

---

to the true vacuum is just one possibility, and we aren't yet certain enough of some of the details of inflation to rule out others. Another really cool possibility is that the transition to the true vacuum occurred in just a "bubble" within a larger universe that has continued to expand exponentially. In this scenario, known as eternal inflation, such bubble universes are continually being formed as the universe in which they are embedded goes on expanding, much like bubbles forming in a glass of seltzer water — except that the amount of seltzer water is increasing exponentially, which rarely happens with glasses of real seltzer water. At least not the kind of seltzer water you find around here. Anyway, in this scenario the universe as a whole is unimaginably vast even by the prodigal standards of modern astronomy, and growing ever faster, all the while giving birth to zillions of bubble universes — so many that everything that could possibly happen would likely have happened in at least one of them. There would be bubble universes in which the whales were saved, bubble universes in which the whales went extinct, bubble universes in which your brothers and sisters were nice to you, and even bubble universes in which the Detroit Lions won the Super Bowl. Because the regions between them would be expanding at far faster than the speed of light, however, these bubble universes could never have any communication or causal contact with each other. But if the parameters of inflation turn out to be such that inflation is eternal, we would know that they are out there. And how freaky would that be?

[77] This is, of course, not to say that our fate is a happy one: depending on the density of the universe, we will eventually either contract under the gravitational pull back into a tiny universe of temperatures so high that all life is extinguished, or we will continue to expand until the universe thins out into a cold, gray mush ever closer to absolute zero temperature — which will also make life impossible. Current measurements indicate that the expansion of the universe is in fact accelerating, making the latter seem the more likely possibility. But either way we're screwed. Or would be, if we were still going to be around; long before then we will have been done in by the national debt.

[78] In addition to the success of inflationary theory in solving some otherwise intractable cosmological problems and explaining various other features of our universe, fine measurements of fluctuations in the 3° background have recently been found to be consistent with inflation.

And the rest, as they say, is history. Once you've got atoms and molecules, forming stars and planets and then making trees and people is relatively easy.

Although we should perhaps note that, while it has a great deal to say about the kinds of matter the universe can have, how that matter can interact, and how therefore the universe evolved from its beginnings, physics is not metaphysics or teleology. Physics can never satisfy our wonder about why is there something rather than nothing. Or why the universe is such a logical place. While from the point of view of the Darwinian evolution of mathematics and logic it makes perfect sense that the universe should be mathematical and logical at the level of three oranges plus four oranges less two oranges being five oranges, what right does this math and logic then have, when pursued at greater depth, to give rise to special and general relativity, quantum field theory, and other abstract physics far beyond effects directly perceivable in our everyday existence?

22.8.1 The Robertson-Walker Metric & Inflation

While working through the details of the universe's expansion and inflationary cosmology would require more math and a greater knowledge of general relativity than we can reasonably get into here, we can, for the benefit of those interested, outline the derivation of these effects and try to give you a sense of how they come about.

Consider a sample of a gas: on a microscopic (atomic) scale, the gas would look very lumpy and uneven — you would see a large void punctuated by a gas molecule here and a gas molecule there — , but on a macroscopic (human) scale the gas looks very continuous and uniform. The distribution of matter in the universe is very similar: on a small scale, it looks very lumpy and uneven, with a large void punctuated here and there by a star, a solar system, or a galaxy, but on a large enough scale the distribution of matter is in fact very uniform and symmetric. A model universe with a uniform distribution of matter should therefore very accurately approximate our actual universe. If one supposes that both the matter within the universe and the pressure due to the motion of that matter are uniform, the metric $g_{\mu\nu}$ that satisfies the gravitational field equations of general relativity turns out to be the Robertson-Walker metric[79]

$$g_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -\dfrac{\bar{R}^2}{1-kr^2} & 0 & 0 \\ 0 & 0 & -\bar{R}^2 r^2 & 0 \\ 0 & 0 & 0 & -\bar{R}^2 r^2 \sin^2\theta \end{pmatrix} \tag{22.51}$$

[79] For more about the metric and the gravitational field equations, see §10.9.1.

The corresponding expression for the proper time interval $d\tau$ is thus

$$c^2\,d\tau^2 = c^2\,dt^2 - \bar{R}^2(t)\left(\frac{dr^2}{1-kr^2} + r^2\,d\theta^2 + r^2\sin^2\theta\,d\varphi^2\right) \tag{22.52}$$

where t is the time coordinate, and r, θ and φ are spherical spatial coordinates. The constant k corresponds to the curvature (that is, the geometry) of the universe and depends on the density of matter within the universe: for densities greater than the critical density, k is positive, corresponding to a closed universe with a spherical geometry that curves back in on itself; for densities less than the critical density, k is negative, corresponding to an open universe that, like a saddle, curves out and away from itself; and when the density of matter equals the critical density, k vanishes, corresponding to a flat universe. The $\bar{R}(t)$ is a function of time obtained by solving the gravitational equations and corresponds to the overall scale or size of the universe.[80] If the time derivative $\dot{\bar{R}}(t)$ of $\bar{R}(t)$ is positive, the universe is expanding; if $\dot{\bar{R}}(t)$ is negative, the universe is contracting.

When a cosmological constant Λ is included, the gravitational field equations are

$$R_{\mu\nu} - \tfrac{1}{2}g_{\mu\nu}R + \frac{1}{c^2}\Lambda g_{\mu\nu} = -\frac{8\pi G}{c^4}T_{\mu\nu}$$

where G is the familiar universal gravitational constant and the energy-momentum tensor $T_{\mu\nu}$ is determined by the distribution of matter. Using the Robertson-Walker metric in the gravitational field equations with an energy-momentum tensor corresponding to a uniform mass density ρ and pressure p, one obtains, after some calculation, a couple of relations governing the evolution of $\bar{R}(t)$ with respect to time:

$$\frac{\ddot{\bar{R}}}{\bar{R}} = -4\pi G\left(\frac{\rho}{3} + \frac{p}{c^2}\right) + \frac{\Lambda}{3} \tag{22.53a}$$

$$\left(\frac{\dot{\bar{R}}}{\bar{R}}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{\bar{R}^2} + \frac{\Lambda}{3} \tag{22.53b}$$

For epochs like the present, which have a vanishing (or perhaps nonzero but minuscule) cosmological constant, so that $\Lambda \approx 0$, eqq. (22.53) reduce to

[80] For those of you worried about the physical dimensions of the various quantities in the Robertson-Walker metric, note that r is not a radial coordinate in the usual sense of a distance: the length dimensions are in $\bar{R}$; both r and k are dimensionless.

Also, just so you know, there is an unfortunate collision in the notation conventionally used: in most books, R stands both for the function $R(t)$ that occurs in the Robertson-Walker metric and for the various forms of the curvature tensor (the Riemann tensor $R^{\lambda}{}_{\mu\nu\kappa}$, the Ricci tensor $R_{\mu\nu}$, and the scalar curvature R). To avoid confusion we have, somewhat unconventionally, put a bar over the former: $\bar{R}(t)$.

the Friedmann equations

$$\frac{\ddot{\bar{R}}}{\bar{R}} = -4\pi G\left(\frac{\rho}{3} + \frac{p}{c^2}\right) \tag{22.54a}$$

$$\left(\frac{\dot{\bar{R}}}{\bar{R}}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{\bar{R}^2} \tag{22.54b}$$

There are then three possibilities, depending on the value of the curvature constant k:

1. If the density of matter is small enough that k is negative (an open universe), then both terms on the right-hand side of eq. (22.54b) are positive, which means that an initially positive $\dot{\bar{R}}/\bar{R}$ will remain positive: the universe will continue expanding forever.

2. If the density of matter is large enough that k is positive (a closed universe), then the two terms on the right-hand side of eq. (22.54b) have opposite signs. As the universe expands, the matter within it is distributed over a larger volume, so that the density ρ of matter decreases, and no matter how large ρ is initially, eventually we will reach a point where the two terms on the right-hand side of eq. (22.54b) will cancel. At this point, $\dot{\bar{R}} = 0$ and the universe will cease expanding. As you can see from eq. (22.54a), ρ and p always make a negative contribution to $\ddot{\bar{R}}$, so that after the expansion ceases, $\dot{\bar{R}}/\bar{R}$ will reverse sign and the universe will start to contract.

3. If the density of matter equals the critical density (a flat universe), then $k = 0$ and the right-hand side of eq. (22.54b) is again always positive: even though ρ and hence the rate of expansion decrease as the universe expands, the expansion continues forever.

Although current astronomical measurements indicate that we are very close to the critical density, because of the difficulty of making the requisite measurements accurately enough, it is not yet known on which side of it we fall. But no matter what we're screwed: if we are at or below the critical density, the universe will continue to expand and thin out into a cold, gray mush ever closer to absolute zero temperature; if we are above the critical density, the universe will eventually cease its expansion and begin to contract back in on itself, with equally catastrophic consequences. You just can't win.

Anyway, the quantity $\dot{\bar{R}}/\bar{R}$, which gives a proportional measure of the rate of expansion, is known as the Hubble constant H:[81]

$$H = \frac{\dot{\bar{R}}}{\bar{R}} \approx 0.71\,\text{km/sec·Mpc} \approx 2.3 \times 10^{-18}\,\text{sec}^{-1}$$

[81] It might strike the alert reader as odd that $\dot{\bar{R}}/\bar{R}$, which is a function of time, is referred

at the present time. If we denote the critical density that corresponds to $k = 0$ by ρ_c, then, using $\rho = \rho_c$, $k = 0$, and $\dot{\bar{R}}/\bar{R} = H$ in eq. (22.54b), we have

$$H^2 = \frac{8\pi G}{3}\rho_c$$

or

$$\rho_c = \frac{3}{8\pi G}H^2 \approx 9.5 \times 10^{-30}\,\mathrm{g/cm^3}$$

which is quite small: the equivalent of roughly just a few hydrogen atoms per cubic meter.

During the epoch of inflation, the cosmological constant Λ dominated the right-hand sides of eqq. (22.53) and the other terms could by comparison be neglected. In this case eqq. (22.53) reduce to

$$\frac{\ddot{\bar{R}}}{\bar{R}} = \tfrac{1}{3}\Lambda \tag{22.55a}$$

$$\left(\frac{\dot{\bar{R}}}{\bar{R}}\right)^2 = \tfrac{1}{3}\Lambda \tag{22.55b}$$

If we separate variables, eq. (22.55b) yields

$$\begin{aligned}
\frac{\dot{\bar{R}}}{\bar{R}} &= \sqrt{\tfrac{1}{3}\Lambda} \\
\frac{1}{\bar{R}}\frac{d\bar{R}}{dt} &= \sqrt{\tfrac{1}{3}\Lambda} \\
\frac{d\bar{R}}{\bar{R}} &= \sqrt{\tfrac{1}{3}\Lambda}\,dt
\end{aligned}$$

Integrating both sides, starting from $\bar{R}(t) = \bar{R}_0$ at $t = 0$,[82] we have

$$\begin{aligned}
\int_{\bar{R}_0}^{\bar{R}} \frac{d\bar{R}}{\bar{R}} &= \int_0^t \sqrt{\tfrac{1}{3}\Lambda}\,dt \\
\ln \bar{R}\Big|_{\bar{R}_0}^{\bar{R}} &= \sqrt{\tfrac{1}{3}\Lambda}\,t
\end{aligned}$$

to as the Hubble *constant*, but *constant* here means spatially constant, not temporally constant.

Also, the values quoted here for the Hubble constant and critical density were obtained from `http://pdg.lbl.gov/2004/reviews/astrorpp.pdf`. A parsec (pc) is a unit of length, $1\ \mathrm{pc} = 3.0856775807 \times 10^{16}$ m.

[82] We are setting up our time coordinate so that $t = 0$ at the start of the inflationary epoch; time zero in the sense of the time at which the universe began is an entirely different issue.

$$\ln\left(\frac{\bar{R}}{\bar{R}_0}\right) = \sqrt{\tfrac{1}{3}\Lambda}\, t$$

$$\bar{R} = \bar{R}_0 \exp\left(\sqrt{\tfrac{1}{3}\Lambda}\, t\right)$$

Thus during the period of inflation the size of the universe was increasing exponentially.

Note in particular that a cosmological constant Λ, unlike the density of matter, does not decrease because of dilution as the universe expands during inflation: Λ is determined by the vacuum value of the inflaton field, and as the universe expands, there is simply more vacuum, throughout which the inflaton field has its vacuum value. It was the tunneling of the inflaton field from the false to the true vacuum that made Λ vanish and brought about the end of the inflationary epoch of our universe.

22.9 Problems

1. Pursue a topic, perhaps in relativity, particle physics, or cosmology, that you find interesting. Research and write a paper on this topic.

   Things to note:

   - This assignment will take the place of the final exam.[83] While not as major as, say, a history term paper, it is a major assignment to which you will need to devote considerable time and energy. Ample homework time will be devoted to this assignment — use it wisely.
   - Although, depending on your topic, you will probably not have the math to actually work out the physics you are researching, your paper should not be simply a book report; it should demonstrate a coherent (if largely qualitative) understanding of the topic.
   - Do not procrastinate. Your first step will be to decide on a possible topic; the next step will be to find material on that topic in the Library, on the Internet, etc. You may find that you need help finding a topic or resources, or, after getting a ways into a topic, that it isn't as suitable or interesting as you at first thought. If you procrastinate, you won't have time to work through these difficulties and pull everything together.

[83] If we assign it, that is; we may do something else.

Appendix A

Projectile Motion with Air Resistance

Projectile motion with velocity-dependent air resistance is a bit beyond what is normally covered in an introductory course, but it is an excellent illustration of some general series and approximation techniques for which you will have frequent use if you go on in physics. In §A.1 we will set up and solve the equations of motion to a first approximation under the assumption that air resistance has only a very small effect on the motion the projectile. In §A.2 we will tackle the exact equations of motion by a more powerful method involving series expansion and successive approximation that, while it still assumes air resistance is in some sense small, allows us to determine the corrections due to air resistance to any degree of accuracy. Finally, in §A.3 we will say a little about solving the equations of motion by numerical integration and show some results so obtained.

For speeds up to 100 mph, air resistance is to a good approximation proportional to the square of a body's velocity v. To include a frictional force proportional to v^2 in projectile motion would, however, make the calculation of quantities like the time of flight and the range very difficult. We will therefore restrict ourselves to the case of an air resistance proportional to v. While such a drag force is somewhat unrealistic, it will make the calculation much more tractable, and the results we will obtain are not unlike those for a frictional force proportional to v^2.

We begin by adding a velocity-proportional term to our projectile-motion force relation, so that the equation of motion $\mathbf{F} = m\mathbf{a}$ becomes

$$m\mathbf{g} - b\mathbf{v} = m\mathbf{a}$$

where as a vector $\mathbf{g}$ is of course directed vertically downward, and where b is a positive constant that measures the strength of the frictional force. By the negative sign on the $b\mathbf{v}$ term, we have taken into account that the force

of air resistance is always opposite to the direction of motion. If we use the dot notation for time derivatives, then, in terms of the position vector $\mathbf{r}$,

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \dot{\mathbf{r}} \qquad\qquad \mathbf{a} = \frac{d^2\mathbf{r}}{dt^2} = \ddot{\mathbf{r}}$$

and the equation of motion is

$$m\mathbf{g} - b\dot{\mathbf{r}} = m\ddot{\mathbf{r}}$$

Solving this for the acceleration $\ddot{\mathbf{r}}$ yields

$$\ddot{\mathbf{r}} = \mathbf{g} - \frac{b}{m}\,\dot{\mathbf{r}}$$

For convenience, so that we don't have to write b/m all the time, we will define

$$\beta = \frac{b}{m}$$

In terms of β our equation for the acceleration of the projectile is

$$\ddot{\mathbf{r}} = \mathbf{g} - \beta\dot{\mathbf{r}} \tag{A.1}$$

Recall now that the components of the position vector $\mathbf{r}$ are simply the coordinates, so that, taking time derivatives, we have

$$\begin{aligned}
\mathbf{r} &= x\,\hat{\mathbf{x}} + y\,\hat{\mathbf{y}} \\
\dot{\mathbf{r}} &= \dot{x}\,\hat{\mathbf{x}} + \dot{y}\,\hat{\mathbf{y}} \\
\ddot{\mathbf{r}} &= \ddot{x}\,\hat{\mathbf{x}} + \ddot{y}\,\hat{\mathbf{y}}
\end{aligned}$$

Orienting our x and y axes in the usual way, with x horizontal and positive y vertically upward, we also have

$$\mathbf{g} = -g\,\hat{\mathbf{y}}$$

The horizontal and vertical components of eq. (A.1) are thus

$$\ddot{x} = -\beta\dot{x} \tag{A.2a}$$

$$\ddot{y} = -g - \beta\dot{y} \tag{A.2b}$$

If air resistance is small, the parameter b, and hence $\beta = b/m$, will be small. But how exactly do we quantify *small*? One way is to recall that there are only two forces acting on the projectile — the gravitational force $m\mathbf{g}$ and the frictional force $-b\mathbf{v}$ — , so that small air resistance means that the magnitude of the frictional force is much less than that of the gravitational force: $bv \ll mg$, or in other words at all times the projectile's velocity v must

be far less than $mg/b = g/\beta$. Another way of quantifying *small* is to note that β has inverse time dimensions: since bv is a force,

$$[b][v] = [\text{force}]$$
$$[b]\frac{[\text{length}]}{[\text{time}]} = \frac{[\text{mass}]\cdot[\text{length}]}{[\text{time}]^2}$$
$$[b] = \frac{[\text{mass}]}{[\text{time}]}$$

so that

$$[\beta] = \frac{[b]}{[\text{mass}]} = [\text{time}]^{-1}$$

Thus small air resistance in some sense means times t such that $\beta t \ll 1$. Why we would want to quantify *small* in this latter way may seem a bit obscure at the moment, but it will become more clear later in the calculation.

A.1 Solution by First Approximation

We will now solve eqq. (A.2) to a first approximation. The idea is that when there is air resistance we can regard x and y as functions not only of time, but also of β:

$$x = x(t,\beta) \qquad\qquad y = y(t,\beta)$$

For small β, we can then do a Taylor expansion of x and y about $\beta = 0$:[1]

$$x = \sum_{n=0}^{\infty}\frac{1}{n!}\beta^n\left[\frac{\partial^n}{\partial\beta^n}x(t,\beta)\right]_{\beta=0} \qquad y = \sum_{n=0}^{\infty}\frac{1}{n!}\beta^n\left[\frac{\partial^n}{\partial\beta^n}y(t,\beta)\right]_{\beta=0}$$

If β is very small, we can to a good approximation neglect terms that go as β^2 and higher powers of β, keeping only those that go as β^0 and β^1:

$$x \approx \Big[x(t,\beta)\Big]_{\beta=0} + \beta\left[\frac{\partial}{\partial\beta}x(t,\beta)\right]_{\beta=0}$$
$$y \approx \Big[y(t,\beta)\Big]_{\beta=0} + \beta\left[\frac{\partial}{\partial\beta}y(t,\beta)\right]_{\beta=0}$$

With the less cumbersome notation

$$\Big[x(t,\beta)\Big]_{\beta=0} = x_{(0)}(t) \qquad\qquad \Big[y(t,\beta)\Big]_{\beta=0} = y_{(0)}(t)$$

[1] We do not need to allow for negative powers of β because the time of flight should of course behave nicely in the limit of no air resistance, that is, as $\beta \to 0$.

$$\left[\frac{\partial}{\partial\beta}x(t,\beta)\right]_{\beta=0} = x_{(1)}(t) \qquad \left[\frac{\partial}{\partial\beta}y(t,\beta)\right]_{\beta=0} = y_{(1)}(t)$$

we may write these relations as

$$x(t) \approx x_{(0)}(t) + \beta x_{(1)}(t) \tag{A.3a}$$
$$y(t) \approx y_{(0)}(t) + \beta y_{(1)}(t) \tag{A.3b}$$

where $x_{(0)}(t)$ and $y_{(0)}(t)$, which correspond to the β^0 terms, are the only contribution to $x(t)$ and $y(t)$ when $\beta = 0$ and should therefore reproduce the usual solutions for the case of projectile motion without air resistance, and where $x_{(1)}(t)$ and $y_{(1)}(t)$, which correspond to the β^1 terms, are the first-order corrections to $x(t)$ and $y(t)$ due to air resistance.

The object of the game from this point is to use the equations of motion to determine $x_{(0)}$, $y_{(0)}$, $x_{(1)}$, and $y_{(1)}$. This will involve integrating the equations of motion from the initial time $t = 0$. Since in the lower limits of these integrations we will need the initial values of $x_{(0)}$, $y_{(0)}$, $x_{(1)}$, and $y_{(1)}$ and of their derivatives $\dot{x}_{(0)}$, $\dot{y}_{(0)}$, $\dot{x}_{(1)}$, and $\dot{y}_{(1)}$, we might as well work out these values now so that we will have them handy when we need them. First of all, we will take $x = y = 0$ at $t = 0$. Since this is an exact relation that holds to all orders, we must have

$$x_{(0)} = y_{(0)} = x_{(1)} = y_{(1)} = 0 \qquad (t = 0)$$

Since air resistance has yet to act at $t = 0$, it will not yet have had any effect on the velocity, so that

$$\dot{x}_{(1)} = \dot{y}_{(1)} = 0 \qquad (t = 0)$$

If we take as the components of the initial velocity $\dot{x}_0 = v_{0x}$ and $\dot{y}_0 = v_{0y}$, we therefore have

$$\begin{aligned} \dot{x}_{(0)} &= v_{0x} \\ \dot{y}_{(0)} &= v_{0y} \end{aligned} \qquad (t = 0)$$

where the 0 subscript, when not in parentheses, has the previously familiar meaning of *initial*, in the sense of *at time* $t = 0$.

Now down to business. Using eq. (A.3a) in eq. (A.2a), we have [2]

$$\ddot{x} = -\beta\dot{x}$$

[2] We could, of course, less painfully solve eqq. (A.2) exactly as far as possible and only then resort to approximation (as we will in §A.2). We could also save ourselves considerable pain by substituting for $x_{(0)}$ and $y_{(0)}$ the solutions that we know, from having previously solved the case of projectile motion without air resistance, they must end up having. But we are going through all this more as an illustration of the general method than simply to get a result for the case of projectile motion with air resistance.

$$\begin{aligned}\ddot{x}_{(0)} + \beta\ddot{x}_{(1)} &= -\beta\Big(\dot{x}_{(0)} + \beta\dot{x}_{(1)}\Big) \\ &= -\beta\dot{x}_{(0)} - \beta^2\dot{x}_{(1)} \end{aligned} \tag{A.4}$$

Since we are working throughout only to first order in β (that is, to order β^1), we can and should neglect any contributions of order β^2 and higher, so we will drop the second term on the right-hand side of eq. (A.4). Moreover, since our relations should hold not just for small β, but for *all* small β as a function of β, we may separately equate terms that go as β^0 and terms that go as β^1. Eq. (A.4) thus gives us two relations:

$$\ddot{x}_{(0)} = 0 \tag{A.5a}$$

$$\ddot{x}_{(1)} = -\dot{x}_{(0)} \tag{A.5b}$$

Eq. (A.5a) is telling us exactly what we expect: that in the absence of air resistance, there is no horizontal acceleration, so that $\dot{x}_{(0)} = \text{const}$. To determine the value of this constant, we have only to note that $\dot{x}_{(0)}$ represents the contribution to the horizontal component v_x of the velocity in the absence of air resistance, so that the constant must be v_{0x}, the initial value of v_x. Thus, if we start from $x = 0$ at $t = 0$,

$$\dot{x}_{(0)} = v_{0x} \tag{A.6a}$$

$$\begin{aligned}\frac{dx_{(0)}}{dt} &= v_{0x} \\ \int_0^{x_{(0)}} dx_{(0)} &= \int_0^t dt\, v_{0x} \\ x_{(0)} &= v_{0x}t \end{aligned} \tag{A.6b}$$

We have now succeeded in solving for $x_{(0)}$, and it does indeed, as expected, reproduce the solution for $x(t)$ for the case of projectile motion without air resistance. Using eq. (A.6b) in eq. (A.5b), we obtain for the first-order correction

$$\ddot{x}_{(1)} = -\dot{x}_{(0)} = -v_{0x}$$

When Integrated twice, this gives

$$\ddot{x}_{(1)} = -v_{0x} \tag{A.7a}$$

$$\begin{aligned}\frac{d\dot{x}_{(1)}}{dt} &= -v_{0x} \\ \int_0^{\dot{x}_{(1)}} d\dot{x}_{(1)} &= -\int_0^t dt\, v_{0x} \\ \dot{x}_{(1)} &= -v_{0x}t \end{aligned}$$

$$\frac{dx_{(1)}}{dt} = -v_{0x}t$$

$$\int_0^{x_{(1)}} dx_{(1)} = -\int_0^t dt\, v_{0x} t$$
$$x_{(1)} = -\tfrac{1}{2} v_{0x} t^2 \tag{A.7b}$$

Putting together eqq. (A.6a) through (A.7b), we now have solutions for x and $v_x = \dot{x}$:

$$\begin{aligned} x &\approx x_{(0)} + \beta x_{(1)} \\ &= v_{0x} t + \beta(-\tfrac{1}{2} v_{0x} t^2) \\ &= v_{0x} t(1 - \tfrac{1}{2}\beta t) \end{aligned} \tag{A.8a}$$
$$\begin{aligned} v_x &\approx \dot{x}_{(0)} + \beta \dot{x}_{(1)} \\ &= v_{0x} + \beta(-v_{0x} t) \\ &= v_{0x}(1 - \beta t) \end{aligned} \tag{A.8b}$$

From these solutions, we can see that, as we would expect, the effect of air resistance is to progressively reduce the horizontal velocity and thus the horizontal distance traveled.

Now we go through the same routine for the y components: using eq. (A.3b) in eq. (A.2b), we have

$$\ddot{y} = -g - \beta \dot{y}$$
$$\ddot{y}_{(0)} + \beta \ddot{y}_{(1)} = -g - \beta\Big(\dot{y}_{(0)} + \beta \dot{y}_{(1)}\Big)$$

Neglecting the terms in β^2 and separately equating the terms in β^0 and β^1 gives

$$\ddot{y}_{(0)} = -g \tag{A.9a}$$
$$\ddot{y}_{(1)} = -\dot{y}_{(0)} \tag{A.9b}$$

Integrating eq. (A.9a) twice gives

$$\ddot{y}_{(0)} = -g$$
$$\frac{d\dot{y}_{(0)}}{dt} = -g$$
$$\int_{v_{0y}}^{\dot{y}_{(0)}} d\dot{y}_{(0)} = -\int_0^t dt\, g$$
$$\dot{y}_{(0)} - v_{0y} = -gt$$
$$\dot{y}_{(0)} = v_{0y} - gt \tag{A.10}$$
$$\frac{dy_{(0)}}{dt} = v_{0y} - gt$$

$$\int_0^{y_{(0)}} dy_{(0)} = \int_0^t dt\,(v_{0y} - gt)$$
$$y_{(0)} = v_{0y}t - \tfrac{1}{2}gt^2 \tag{A.11}$$

Our solutions for $y_{(0)}$ and $\dot{y}_{(0)}$ do indeed, as we would expect, reproduce our results for y and v_y for the case of projectile motion without air resistance.

Using eq. (A.10) in eq. (A.9b) and integrating twice, we also obtain

$$\ddot{y}_{(1)} = -\dot{y}_{(0)}$$
$$\frac{d\dot{y}_{(1)}}{dt} = -(v_{0y} - gt)$$
$$\int_0^{\dot{y}_{(1)}} d\dot{y}_{(1)} = -\int_0^t dt\,(v_{0y} - gt)$$
$$\dot{y}_{(1)} = -v_{0y}t + \tfrac{1}{2}gt^2 \tag{A.12}$$
$$\frac{dy_{(1)}}{dt} = -v_{0y}t + \tfrac{1}{2}gt^2$$
$$\int_0^{y1} dy_{(1)} = \int dt\,(-v_{0y}t + \tfrac{1}{2}gt^2)$$
$$y_{(1)} = -\tfrac{1}{2}v_{0y}t^2 + \tfrac{1}{6}gt^3 \tag{A.13}$$

Putting together eqq. (A.10) through (A.13), we have solutions for y and $v_y = \dot{y}$:

$$\begin{aligned} y &\approx y_{(0)} + \beta y_{(1)} \\ &= (v_{0y}t - \tfrac{1}{2}gt^2) + \beta(-\tfrac{1}{2}v_{0y}t^2 + \tfrac{1}{6}gt^3) \\ &= v_{0y}t(1 - \tfrac{1}{2}\beta t) - \tfrac{1}{2}gt^2(1 - \tfrac{1}{3}\beta t) \end{aligned} \tag{A.14a}$$
$$\begin{aligned} v_y &\approx \dot{y}_{(0)} + \beta\dot{y}_{(1)} \\ &= (v_{0y} - gt) + \beta(-v_{0y}t + \tfrac{1}{2}gt^2) \\ &= v_{0x}(1 - \beta t) - gt(1 - \tfrac{1}{2}\beta t) \end{aligned} \tag{A.14b}$$

From these solutions, we can see that the effect of air resistance is to progressively reduce the initial vertical velocity and also the downward gravitational contribution $-gt$ to the vertical velocity, with corresponding effects on the vertical distance traveled.

To prolong your agony, we will now work out the range of the projectile, limiting your suffering only by assuming that the projectile is fired over level ground. As in the case of projectile motion without air resistance, we determine the range by solving the vertical-displacement equation for the time of flight and then using that time in the horizontal-displacement equation. Since at the far end of the trajectory y returns to its initial value, which we

took to be $y_0 = 0$, we set $y = 0$ in eq. (A.14a):

$$0 = v_{0y}(1 - \tfrac{1}{2}\beta t) - \tfrac{1}{2}gt(1 - \tfrac{1}{3}\beta t)$$

$t = 0$ is obviously not the solution we want; dividing through by t leaves

$$0 = v_{0y}(1 - \tfrac{1}{2}\beta t) - \tfrac{1}{2}gt(1 - \tfrac{1}{3}\beta t) \tag{A.15}$$

This is a quadratic equation that we could solve exactly for t — except that it is based on a solution for y that is valid only to order β^1 to begin with. We therefore want to solve eq. (A.15) for t only to first order in β, for which purpose we will express t in the same way that we did x and y:

$$t \approx t_{(0)} + \beta t_{(1)}$$

Using this approximation for t in eq. (A.15) gives, when we expand and drop terms in β^2 and higher,

$$\begin{aligned}
0 &= v_{0y}\Big(1 - \tfrac{1}{2}\beta(t_{(0)} + \beta t_{(1)})\Big) - \tfrac{1}{2}g(t_{(0)} + \beta t_{(1)})\Big(1 - \tfrac{1}{3}\beta(t_{(0)} + \beta t_{(1)})\Big) \\
&= v_{0y}(1 - \tfrac{1}{2}\beta t_{(0)}) - \tfrac{1}{2}g(t_{(0)} + \beta t_{(1)})(1 - \tfrac{1}{3}\beta t_{(0)}) \\
&= v_{0y}(1 - \tfrac{1}{2}\beta t_{(0)}) - \tfrac{1}{2}g(t_{(0)} - \tfrac{1}{3}\beta t_{(0)}^2 + \beta t_{(1)}) \\
&= (v_{0y} - \tfrac{1}{2}gt_{(0)}) + \beta(-\tfrac{1}{2}v_{0y}t_{(0)} + \tfrac{1}{6}gt_{(0)}^2 - \tfrac{1}{2}gt_{(1)})
\end{aligned}$$

Again separately equating the β^0 and β^1 terms, we obtain

$$\begin{aligned}
0 &= v_{0y} - \tfrac{1}{2}gt_{(0)} \\
t_{(0)} &= \frac{2v_{0y}}{g} \qquad \text{(A.16a)} \\
0 &= -\tfrac{1}{2}v_{0y}t_{(0)} + \tfrac{1}{6}gt_{(0)}^2 - \tfrac{1}{2}gt_{(1)} \\
&= -\tfrac{1}{2}v_{0y}\left(\frac{2v_{0y}}{g}\right) + \tfrac{1}{6}g\left(\frac{2v_{0y}}{g}\right)^2 - \tfrac{1}{2}gt_{(1)} \\
&= -\frac{v_{0y}^2}{3g} - \tfrac{1}{2}gt_{(1)} \\
t_{(1)} &= -\frac{2v_{0y}^2}{3g^2} \qquad \text{(A.16b)}
\end{aligned}$$

The result for $t_{(0)}$ is the same as the time of flight we previously worked out for the case of projectile motion without air resistance. Also note that the correction term $t_{(1)}$ is always negative: air resistance always decreases the time of flight.

Putting (A.16a) and (A.16b) together, we have for the time of flight

$$t \approx t_{(0)} + \beta t_{(1)} = \frac{2v_{0y}}{g} + \beta\left(-\frac{2v_{0y}^2}{3g^2}\right) = \frac{2v_{0y}}{g}\left(1 - \frac{\beta v_{0y}}{3g}\right)$$

Using this result for t in the horizontal-displacement relation (A.8a), and as usual dropping terms of order β^2, we obtain for the range

$$\begin{aligned}
R &= v_{0x}t(1 - \tfrac{1}{2}\beta t) \\
&= v_{0x}\left(\frac{2v_{0y}}{g}\left(1 - \frac{\beta v_{0y}}{3g}\right)\right)\left(1 - \tfrac{1}{2}\beta\left(\frac{2v_{0y}}{g}\left(1 - \frac{\beta v_{0y}}{3g}\right)\right)\right) \\
&= v_{0x}\frac{2v_{0y}}{g}\left(1 - \frac{\beta v_{0y}}{3g}\right)\left(1 - \tfrac{1}{2}\beta\left(\frac{2v_{0y}}{g}\right)\right) \\
&= v_{0x}\frac{2v_{0y}}{g}\left(1 - \frac{\beta v_{0y}}{3g}\right)\left(1 - \frac{\beta v_{0y}}{g}\right) \\
&= v_{0x}\frac{2v_{0y}}{g}\left(1 - \frac{\beta v_{0y}}{3g} - \frac{\beta v_{0y}}{g}\right) \\
&= \frac{2v_{0x}v_{0y}}{g}\left(1 - \frac{4}{3}\frac{\beta v_{0y}}{g}\right)
\end{aligned} \tag{A.17}$$

The $2v_{0x}v_{0y}/g$ part of this is just our previous result for the range in the absence of air resistance. As you can see, the effect of air resistance is, as we would expect, always to decrease the range.

We could call it a day at this point, but to prolong your agony still further, we will now determine the angle of launch that will maximize the range of the projectile. First, we need to rewrite eq. (A.17) in terms of the launch angle θ with the horizontal: with

$$v_{0x} = v_0\cos\theta \qquad\qquad v_{0y} = v_0\sin\theta$$

eq. (A.17) becomes

$$\begin{aligned}
R &= \frac{2(v_0\cos\theta)(v_0\sin\theta)}{g}\left(1 - \frac{4}{3}\frac{\beta(v_0\sin\theta)}{g}\right) \\
&= \frac{v_0^2(2\sin\theta\cos\theta)}{g}\left(1 - \frac{4}{3}\frac{\beta v_0\sin\theta}{g}\right) \\
&= \frac{v_0^2\sin 2\theta}{g}\left(1 - \frac{4}{3}\frac{\beta v_0\sin\theta}{g}\right)
\end{aligned}$$

Taking the θ derivative of this to maximize R then gives

$$0 = \frac{2v_0^2\cos 2\theta}{g}\left(1 - \frac{4}{3}\frac{\beta v_0\sin\theta}{g}\right) + \frac{v_0^2\sin 2\theta}{g}\left(-\frac{4}{3}\frac{\beta v_0\cos\theta}{g}\right)$$

which, if we divide out a $2v_0^2/g$ and tidy up a bit, "reduces" to

$$\begin{aligned}
0 &= \cos 2\theta\left(1 - \frac{4}{3}\frac{\beta v_0 \sin\theta}{g}\right) + \tfrac{1}{2}\sin 2\theta\left(-\frac{4}{3}\frac{\beta v_0 \cos\theta}{g}\right) \\
&= \cos 2\theta - \frac{4\beta v_0}{3g}\Big(\cos 2\theta \sin\theta + \tfrac{1}{2}\sin 2\theta \cos\theta\Big) \\
&= \cos 2\theta - \frac{4\beta v_0}{3g}\Big((2\cos^2\theta - 1)\sin\theta + \tfrac{1}{2}(2\sin\theta\cos\theta)\cos\theta\Big) \\
&= \cos 2\theta - \frac{4\beta v_0}{3g}(3\cos^2\theta\sin\theta - \sin\theta) \\
&= \cos 2\theta - \frac{4\beta v_0}{3g}(3\cos^2\theta - 1)\sin\theta \qquad \text{(A.18)}
\end{aligned}$$

Following the usual prescription, to solve for θ we first express it to first order as

$$\theta \approx \theta_{(0)} + \beta\theta_{(1)}$$

so that eq. (A.18) becomes

$$0 = \cos 2(\theta_{(0)} + \beta\theta_{(1)}) - \frac{4\beta v_0}{3g}\Big(3\cos^2(\theta_{(0)} + \beta\theta_{(1)}) - 1\Big)\sin(\theta_{(0)} + \beta\theta_{(1)}) \quad \text{(A.19)}$$

We now expand this, dropping terms of order β^2 and higher. In the terms with the overall factor of $4\beta v_0/3g$, we can drop the $\theta_{(1)}$'s: these would contribute corrections of order β^1 and higher to terms that already go as β^1 because of the factor of $4\beta v_0/3g$. We need expand only the $\cos 2(\theta_{(0)} + \beta\theta_{(1)})$, which is most easily accomplished by doing the Taylor series out to first order and using the chain rule to evaluate the derivative:

$$\begin{aligned}
\cos 2(\theta_{(0)} + \beta\theta_{(1)}) &\approx \Big[\cos 2(\theta_{(0)} + \beta\theta_{(1)})\Big]_{\beta=0} + \beta\left[\frac{\partial}{\partial\beta}\cos 2(\theta_{(0)} + \beta\theta_{(1)})\right]_{\beta=0} \\
&= \cos 2\theta_{(0)} + \beta(-2\sin 2\theta_{(0)} \cdot \theta_{(1)}) \\
&= \cos 2\theta_{(0)} - 2\beta\theta_{(1)}\sin 2\theta_{(0)}
\end{aligned}$$

Using this in eq. (A.19), we have

$$0 = \cos 2\theta_{(0)} - 2\beta\theta_{(1)}\sin 2\theta_{(0)} - \frac{4\beta v_0}{3g}(3\cos^2\theta_{(0)} - 1)\sin\theta_{(0)}$$

When we separately equate terms of order β^0 and β^1, this yields

$$\begin{aligned}
0 &= \cos 2\theta_{(0)} \\
\theta_{(0)} &= \tfrac{1}{2}\cos^{-1} 0 = \tfrac{1}{2}\,\tfrac{\pi}{2} = \tfrac{\pi}{4}
\end{aligned}$$

$$\begin{aligned}
0 &= -2\theta_{(1)} \sin 2\theta_{(0)} - \frac{4v_0}{3g}(3\cos^2\theta_{(0)} - 1)\sin\theta_{(0)} \\
&= -2\theta_{(1)} \sin 2\frac{\pi}{4} - \frac{4v_0}{3g}(3\cos^2\frac{\pi}{4} - 1)\sin\frac{\pi}{4} \\
&= -2\theta_{(1)}(1) - \frac{4v_0}{3g}\left(3\left(\frac{1}{\sqrt{2}}\right)^2 - 1\right)\frac{1}{\sqrt{2}} \\
&= -2\theta_{(1)} - \frac{2v_0}{3g\sqrt{2}} \\
\theta_{(1)} &= -\frac{v_0}{3g\sqrt{2}} \\
\theta &\approx \theta_{(0)} + \beta\theta_{(1)} = \frac{\pi}{4} - \frac{\beta v_0}{3g\sqrt{2}}
\end{aligned}$$

This reproduces the expected $\frac{\pi}{4}$ in the absence of air resistance. Also note that with air resistance the range is maximized by a shallower rather than a steeper launch angle.

A.2 Solution by Series Expansion

In this section, we will solve the equations of motion (A.2) by a series expansion and successive approximation that, while it still assumes β is small, does not assume β is so small that terms of order β^2 and higher are negligible; rather, this method allows us to determine the effects of air resistance to any order in β and thus any degree of accuracy.

We begin by returning to eq. (A.2a),

$$\ddot{x} = -\beta\dot{x}$$

which we now solve exactly by separating variables and integrating:

$$\begin{aligned}
\frac{d\dot{x}}{dt} &= -\beta\dot{x} \\
\int_{\dot{x}_0}^{\dot{x}} \frac{d\dot{x}}{\dot{x}} &= -\int_0^t \beta\, dt \\
\ln\left(\frac{\dot{x}}{\dot{x}_0}\right) &= -\beta t \\
\dot{x} &= \dot{x}_0\, e^{-\beta t}
\end{aligned}$$

where $\dot{x}_0$ is the initial x component of the velocity. In other words, $\dot{x}_0 = v_{0x}$, the value of v_x at time $t = 0$:

$$\dot{x} = v_{0x}\, e^{-\beta t} \tag{A.20a}$$

Integrating this once more with respect to time, we have, assuming that the projectile is launched from $x = 0$,

$$\begin{aligned}
\frac{dx}{dt} &= v_{0x}\, e^{-\beta t} \\
\int_0^x dx &= \int_0^t dt\, v_{0x}\, e^{-\beta t} \\
x &= -\frac{v_{0x}}{\beta}\, e^{-\beta t}\bigg|_0^t \\
&= \frac{v_{0x}}{\beta}\Big(1 - e^{-\beta t}\Big) \qquad \text{(A.20b)}
\end{aligned}$$

Having dispensed with eq. (A.2a), we turn to eq. (A.2b):

$$\ddot{y} = -g - \beta\dot{y}$$

Again separating variables and integrating, and using $\dot{y}_0 = v_{0y}$, we obtain

$$\begin{aligned}
\frac{d\dot{y}}{dt} &= -g - \beta\dot{y} \\
&= -\beta\left(\dot{y} + \frac{g}{\beta}\right) \\
\int_{v_{0y}}^{\dot{y}} \frac{d\dot{y}}{\dot{y} + g/\beta} &= -\int_0^t \beta\, dt \\
\ln\left(\frac{\dot{y} + g/\beta}{v_{0y} + g/\beta}\right) &= -\beta t
\end{aligned}$$

which, when solved for $\dot{y}$, yields

$$\dot{y} = \left(v_{0y} + \frac{g}{\beta}\right)\, e^{-\beta t} - \frac{g}{\beta} \qquad \text{(A.21a)}$$

Integrating this once more with respect to time, and assuming that the projectile is launched from $y = 0$, we have

$$\begin{aligned}
\frac{dy}{dt} &= \left(v_{0y} + \frac{g}{\beta}\right)\, e^{-\beta t} - \frac{g}{\beta} \\
\int_0^y dy &= \int_0^t dt \left(\left(v_{0y} + \frac{g}{\beta}\right)\, e^{-\beta t} - \frac{g}{\beta}\right) \\
y &= -\frac{1}{\beta}\left(v_{0y} + \frac{g}{\beta}\right)\, e^{-\beta t}\bigg|_0^t - \frac{g}{\beta}\, t \\
&= \frac{1}{\beta}\left(v_{0y} + \frac{g}{\beta}\right)\Big(1 - e^{-\beta t}\Big) - \frac{g}{\beta}\, t \qquad \text{(A.21b)}
\end{aligned}$$

When the projectile is fired over level ground, the range is the horizontal displacement x at the nonzero time for which y returns to the value $y = 0$. Setting $y = 0$ in eq. (A.21b) yields

$$0 = \frac{1}{\beta}\left(v_{0y} + \frac{g}{\beta}\right)\left(1 - e^{-\beta t}\right) - \frac{g}{\beta}\,t$$

which, if we multiply both sides by β^2, simplifies to

$$0 = (\beta v_{0y} + g)\Big(1 - e^{-\beta t}\Big) - g\beta t \tag{A.22}$$

This is a transcendental equation for t that has no solution in closed form. But if β is small, we can do an expansion of t in powers of β: [3]

$$t = t_{(0)} + \beta t_{(1)} + \beta^2 t_{(2)} + \beta^3 t_{(3)} + \cdots = \sum_{m=0}^{\infty} t_{(m)}\beta^m \tag{A.23}$$

where the $t_{(m)}$ are as yet unknown coefficients. To deal with eq. (A.22) as a power series in β, we will want to do a Taylor expansion of the factor $1 - e^{-\beta t}$:

$$\begin{aligned} 1 - e^{-\beta t} &= 1 - \sum_{n=0}^{\infty} \frac{1}{n!}(-\beta t)^n \\ &= -\sum_{n=1}^{\infty} \frac{1}{n!}(-\beta t)^n \qquad \text{(A.24a)} \\ &= -\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^{n+1} \\ &= \beta t \sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n \qquad \text{(A.24b)} \end{aligned}$$

then eq. (A.22) becomes

$$0 = (\beta v_{0y} + g)\beta t \sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n - g\beta t$$

which, when we divide out an overall βt, reduces to

$$0 = (\beta v_{0y} + g) \sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n - g \tag{A.25}$$

The "trick" is now to use eq. (A.25) to solve for the coefficients $t_{(m)}$ in the series expansion (A.23) for the time of flight t, starting with $t_{(0)}$, then getting $t_{(1)}$, then $t_{(2)}$, and so on, to as fine a degree of approximation as necessary.

[3] We do not need to include negative powers of β because the time of flight should of course behave nicely in the limit of no air resistance, that is, as $\beta \to 0$.

And the key to accomplishing this is to note that eq. (A.25), although it assumes β is small, must hold as a function of β for *all* β within this domain and therefore hold separately for each power of β in the series expansion.

For convenience, we will use the following notation: when brackets around a quantity are followed by a subscript, the subscript indicates the power of β that is to be extracted from the quantity within the brackets. Thus $[z]_1$ means to keep in z only the terms that have a β^1. Note that the terms in eq. (A.25) from the summation consist of several factors that each involve nonnegative integer powers of β; when we want to extract the β^n contribution from such terms, we must consider all the various combinations of powers of β among the factors that will add up to n. For example, suppose we have three factors A, B, and C. Since the only combination of powers of β from A, B, and C that adds up to zero is $0+0+0$,

$$[ABC]_0 = [A]_0\ [B]_0\ [C]_0$$

But when we want the β^1 contribution from ABC, the factor of β could come from any one of the three factors, so that

$$[ABC]_1 = [A]_1\ [B]_0\ [C]_0 + [A]_0\ [B]_1\ [C]_0 + [A]_0\ [B]_0\ [C]_1$$

And to extract the β^2 contribution, we must consider the six combinations

$$2+0+0 \quad 0+2+0 \quad 0+0+2 \quad 1+1+0 \quad 1+0+1 \quad 0+1+1$$

so that

$$\begin{aligned}[ABC]_2 = {} & [A]_2\ [B]_0\ [C]_0 + [A]_0\ [B]_2\ [C]_0 + [A]_0\ [B]_0\ [C]_2 \\ & + [A]_1\ [B]_1\ [C]_0 + [A]_1\ [B]_0\ [C]_1 + [A]_0\ [B]_1\ [C]_1\end{aligned}$$

Okay, now down to business: The lowest-order terms in eq. (A.25) are those that go as β^0:

$$\begin{aligned}0 &= [\beta v_{0y} + g]_0 \left[\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_0 - [g]_0 \\ &= (g)(1) - (g) \\ &= 0\end{aligned}$$

which is certainly true, but not much help.

The next-to-lowest-order terms in eq. (A.25) are those that go as β^1:

$$0 = \left[(\beta v_{0y} + g)\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_1 - [g]_1 \tag{A.26}$$

Since g is purely of order β^0, $[g]_1 = 0$. Expanding the β^1 contribution of the rest of eq. (A.26) as in our ABC example above, and remembering that t here actually represents the series expansion (A.23), we have

$$\begin{aligned}
0 &= [(\beta v_{0y} + g)]_1 \left[\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_0 \\
&\qquad + [(\beta v_{0y} + g)]_0 \left[\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_1 \\
&= (\beta v_{0y})(1) + (g)\left[\frac{1}{2!}(-\beta t)\right]_1 \\
&= \beta v_{0y} + g\left(-\tfrac{1}{2}\beta [t]_0\right) \\
&= \beta v_{0y} + g\left(-\tfrac{1}{2}\beta t_{(0)}\right) \\
&= \beta(v_{0y} - \tfrac{1}{2} g t_{(0)})
\end{aligned} \tag{A.27}$$

which, as we would expect, reproduces the familiar result for time of flight in the absence of air resistance:

$$t_{(0)} = \frac{2v_{0y}}{g} \tag{A.28}$$

To see the leading-order effects of air resistance, we have to look at the β^2 terms in eq. (A.25). With $[g]_2 = 0$, these are

$$\begin{aligned}
0 &= \left[(\beta v_{0y} + g)\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_2 \\
&= [\beta v_{0y} + g]_2 \left[\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_0 \\
&\qquad + [\beta v_{0y} + g]_0 \left[\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_2 \\
&\qquad + [\beta v_{0y} + g]_1 \left[\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_1 \\
&= (0)(1) + (g)\left[\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_2 + (\beta v_{0y})\left(-\tfrac{1}{2}\beta t_{(0)}\right)
\end{aligned} \tag{A.29}$$

where we have in the last step made use of our result (A.27):

$$\left[\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_1 = -\tfrac{1}{2}\beta t_{(0)}$$

To deal with

$$\left[\sum_{n=0}^{\infty} \frac{1}{(n+1)!}(-\beta t)^n\right]_2$$

we need again remember that t here actually represents the series expansion (A.23), so that

$$\begin{aligned}\left[\sum_{n=0}^{\infty}\frac{1}{(n+1)!}(-\beta t)^n\right]_2 &= \frac{1}{2!}[-\beta t]_2 + \frac{1}{3!}\left[(-\beta t)^2\right]_2 \\ &= \tfrac{1}{2}(-\beta[t]_1) + \tfrac{1}{6}\Big(\beta^2[t]_0^2\Big) \\ &= \tfrac{1}{2}\Big(-\beta(\beta t_{(1)})\Big) + \tfrac{1}{6}\Big(\beta^2 t_{(0)}^2\Big) \\ &= \beta^2(-\tfrac{1}{2}t_{(1)} + \tfrac{1}{6}t_{(0)}^2)\end{aligned}$$

Using this result in eq. (A.29), we arrive at

$$\begin{aligned}0 &= (0)(1) + (g)\Big(\beta^2(-\tfrac{1}{2}t_{(1)} + \tfrac{1}{6}t_{(0)}^2)\Big) + (\beta v_{0y})\Big(-\tfrac{1}{2}\beta t_{(0)}\Big) \\ &= \beta^2(-\tfrac{1}{2}gt_{(1)} + \tfrac{1}{6}gt_{(0)}^2 - \tfrac{1}{2}v_{0y}t_{(0)})\end{aligned}$$

which, if we use $t_{(0)} = 2v_{0y}/g$ from eq. (A.28), becomes

$$\begin{aligned}0 &= \beta^2\left(-\tfrac{1}{2}gt_{(1)} + \tfrac{1}{6}g\left(\frac{2v_{0y}}{g}\right)^2 - \tfrac{1}{2}v_{0y}\left(\frac{2v_{0y}}{g}\right)\right) \\ &= \beta^2\left(-\tfrac{1}{2}gt_{(1)} - \frac{v_{0y}^2}{3g}\right)\end{aligned}$$

Solving for $t_{(1)}$, we arrive at

$$t_{(1)} = -\frac{2v_{0y}^2}{3g^2} \tag{A.30}$$

in agreement with our result (A.16b) of the previous section.

Although we will not pursue the calculation further than this first-order result for the time of flight t, the method for doing so should now be clear: The β^1 terms in eq. (A.25) involved only $t_{(0)}$, for which we then solved. The β^2 terms in eq. (A.25) involved $t_{(0)}$ and $t_{(1)}$, which, since we had already solved at the previous order for $t_{(0)}$, allowed us to solve for $t_{(1)}$. The β^3 terms in eq. (A.25) would involve $t_{(0)}$, $t_{(1)}$, and $t_{(2)}$, which, since we would already have results for $t_{(0)}$ and $t_{(1)}$, would allow us to solve for $t_{(2)}$. And so on. A solution for t can in this way be obtained to any desired degree of accuracy, assuming only that β is small enough that our series expansion (A.23) converges. And this same method could be applied to the calculation of other quantities, such as the range and the angle that maximizes the range.

Putting together (A.28) and (A.30), our result for the time of flight, to first order in the air-resistance parameter β, is

$$t \approx t_{(0)} + \beta t_{(1)} = \frac{2v_{0y}}{g} + \beta\left(-\frac{2v_{0y}^2}{3g^2}\right) = \frac{2v_{0y}}{g}\left(1 - \frac{\beta v_{0y}}{3g}\right) \tag{A.31}$$

To see out to the leading-order effect of air resistance on the range, we go back to eq. (A.20b) and, using eq. (A.24a), keep terms that go as β^0 or β^1:

$$\begin{aligned}
R &= \left[\frac{v_{0x}}{\beta}\Big(1-e^{-\beta t}\Big)\right]_0 + \left[\frac{v_{0x}}{\beta}\Big(1-e^{-\beta t}\Big)\right]_1 \\
&= \left[\frac{v_{0x}}{\beta}\left(-\sum_{n=1}^{\infty}\frac{1}{n!}(-\beta t)^n\right)\right]_0 + \left[\frac{v_{0x}}{\beta}\left(-\sum_{n=1}^{\infty}\frac{1}{n!}(-\beta t)^n\right)\right]_1 \\
&= \frac{v_{0x}}{\beta}\left[-\sum_{n=1}^{\infty}\frac{1}{n!}(-\beta t)^n\right]_1 + \frac{v_{0x}}{\beta}\left[-\sum_{n=1}^{\infty}\frac{1}{n!}(-\beta t)^n\right]_2 \\
&= \frac{v_{0x}}{\beta}\Big(-\frac{1}{1!}(-\beta[t]_0)\Big) + \frac{v_{0x}}{\beta}\Big(-\frac{1}{1!}(-\beta[t]_1) - \frac{1}{2!}(-\beta[t]_0)^2\Big) \\
&= \frac{v_{0x}}{\beta}\Big(-\big(-\beta t_{(0)}\big)\Big) + \frac{v_{0x}}{\beta}\Big(-\big(-\beta(\beta t_{(1)})\big) - \tfrac{1}{2}\big(-\beta t_{(0)}\big)^2\Big) \\
&= v_{0x}\Big(t_{(0)} + \beta t_{(1)} - \tfrac{1}{2}\beta t_{(0)}^2\Big)
\end{aligned}$$

Using eqq. (A.28) and (A.31) in this, we arrive at

$$\begin{aligned}
R &= v_{0x}\left(\frac{2v_{0y}}{g} + \beta\left(-\frac{2v_{0y}^2}{3g^2}\right) - \tfrac{1}{2}\beta\left(\frac{2v_{0y}}{g}\right)^2\right) \\
&= \frac{2v_{0x}v_{0y}}{g}\left(1 - \frac{4}{3}\frac{\beta v_{0y}}{g}\right)
\end{aligned}$$

in agreement with our result (A.17) of the previous section. As you can see, the effect of air resistance is, as expected, always to decrease the range.

The determination of the first-order correction to the launch angle that maximizes this range is the same as in the previous section, so we will simply refer you back to p.1121ff.

A.3 Solution by Numerical Integration

In the days before computers, numerical integration of differential equations had to be done by hand and was, as you can imagine, a huge pain. Or perhaps you can't imagine, having been born into the decadent generation of the calculator and computer. Back when we were in school, we had to learn how to interpolate log tables. And at Los Alamos during the Manhattan Project, they needed teams of people working long days, each person responsible for myriad repetitions of one arithmetic step in a huge calculation: he or she would be handed the result of the previous step on an index card, perform his or her arithmetic step, and then pass the card along to the person handling

the following step. In those days, *calculator* meant a person, not a little plastic box. You kids don't know how easy you have it these days. Pshaw!

Anyway, if you know the numerical values of the various parameters in the equation of motion, these days you can use an application like Octave [4] to integrate the equations numerically and obtain as accurate a result as desired, without even having to worry about whether air resistance is small or not. Physicists of course prefer the more general results obtainable by analytical methods and resort to numerical techniques only on those occasions when analytical methods prove intractable. Engineers find themselves more frequently resorting to numerical methods, because there are generally far too many degrees of freedom in practical engineering problems for analytical solution to be feasible.

While the use of such applications is beyond the scope of our studies, we will, just so you will be better prepared when the time comes for you to delve into them, make note of one general feature: they can only integrate first-order differential equations. That is, they can, for example, integrate equations like

$$\frac{dy}{dx} + f(x)\,y = g(x) \tag{A.32}$$

or even

$$\left(\frac{dy}{dx}\right)^2 + f(x)\,\sqrt{y} = g(x)$$

but not

$$\frac{d^2y}{dx^2} + f(x)\,y = g(x)$$

This restriction, which might at first glance seem severe, is a consequence of the way one integrates numerically: for our example (A.32) above, starting at some initial point (x, y), the values of $f(x)$ and $g(x)$ would be calculated, and then the change dy in the value of y for some suitably small change dx in the value of x would be determined by

$$dy = \Big(g(x) - f(x)\,y\Big)\,dx$$

This would bring us to the point $(x + dx, y + dy)$, at which the calculation would be repeated, and so on. Applications that do numerical calculations of course use very sophisticated methods to maximize efficiency and minimize error, but this is the gist of how they work.

It is, however, easy to circumvent the restriction that the equation must be first order and extend this algorithm to the solution of higher-order differential equations: one simply lowers the order of the equation by increasing

[4] An interesting bit of trivia: people tend to think that Octave is named after the musical term, when it's actually the (Italian) name of a former professor of the program's author who had a talent for doing numerical calculations in his head.

the number of variables. If, for example, you want to integrate

$$\frac{d^2y}{dx^2} = f(x)$$

you simply define, in addition to y, a new variable

$$\tilde{y} = \frac{dy}{dx}$$

The single second-order equation $d^2y/dx^2 = f(x)$ then becomes a system of two first-order equations:

$$\frac{dy}{dx} = \tilde{y} \qquad\qquad \frac{d\tilde{y}}{dx} = f(x)$$

These equations may then be integrated numerically to obtain $\tilde{y}(x)$ and $y(x)$.

In our case, we have the equations of motion (A.2):

$$\ddot{x} = \beta\dot{x} \qquad\qquad \ddot{y} = -g - \beta\dot{y}$$

We can make these into a system of first-order equations by introducing $\tilde{x} = \dot{x}$ and $\tilde{y} = \dot{y}$:

$$\begin{aligned} \dot{x} &= \tilde{x} & \dot{\tilde{x}} &= \beta\dot{x} \\ \dot{y} &= \tilde{y} & \dot{\tilde{y}} &= -g - \beta\dot{y} \end{aligned}$$

Numerical integration then yields results for $x(t)$, $y(t)$, $\tilde{x}(t)$, and $\tilde{y}(t)$.

Fig. (A.1) shows the result of such a numerical integration for the mythical set of values

$$g = 10\,\text{m/s}^2 \qquad \beta = 0.1\,\text{sec}^{-1} \qquad v_{0x} = 30\,\text{m/s} \qquad v_{0y} = 15\,\text{m/s}$$

The black curve is the parabolic trajectory we would have had without air resistance, the red curve the trajectory with air resistance. In this case we expect, from eq. (A.17),

$$\begin{aligned} R &= \frac{2v_{0x}v_{0y}}{g}\left(1 - \frac{4}{3}\frac{\beta v_{0y}}{g}\right) \\ &= \frac{2(15)(30)}{10}\left(1 - \frac{4}{3}\frac{(0.1)(15)}{10}\right) \\ &= 90(1 - 0.2) \\ &= 72\,\text{m} \end{aligned}$$

versus the 90 m range that we would have without air resistance. The actual value of the range obtained from the numerical integration is 74.7 m, which is

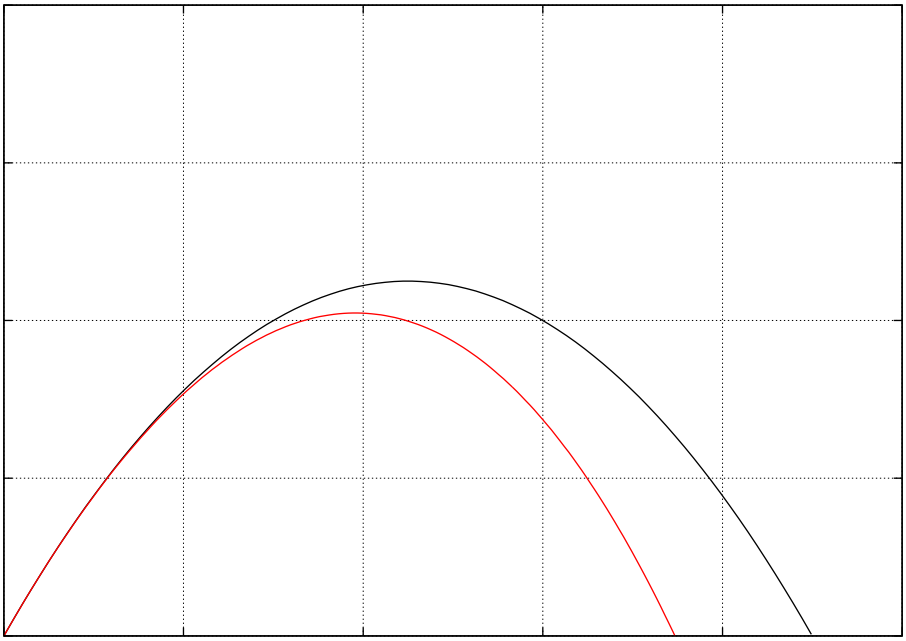


Figure A.1: Air Resistance Proportional to Velocity

well within what we expect to be the limits of the accuracy of our first-order approximation: since the first-order correction to the range was of magnitude $90(0.2)$, we expect that, very roughly, the β^2 correction to the range would be of magnitude $90(0.2^2) = 3.6\,\mathrm{m}$, so that the margin of error on our first-order result of 72 m is very roughly ± 3.6 m. Woohoo!

Fig. (A.2) shows the result of a numerical integration for the more realistic case of an air resistance proportional to the square of the velocity, though with a considerably smaller value of β.[5] Even more realistically, fig. (A.3) shows the effect of an air resistance proportional to the square of the velocity on a 30-06 round, using actual values for the various parameters.[6] In both figures the black curves are the parabolic trajectories we would have had without air resistance, the red curves the trajectory with air resistance. In the case of the 30-06 round, note the general feature that, in addition to not going as high or as far, the shape of the descent is markedly compressed horizontally relative to that of a parabola.

[5] The β's for the cases of air resistance proportional to the velocity and to the square of the velocity have of course different physical dimensions and cannot be compared directly; we mean that relative to the other parameters in the calculation the air resistance is smaller.

[6] Of course, in this case the speed of the 30-06 round — initially 810 m/s — is so high that the approximation that air resistance proportional to the square of the velocity is no longer valid. But except for being totally unrealistic in that respect, the calculation is realistic. ☺

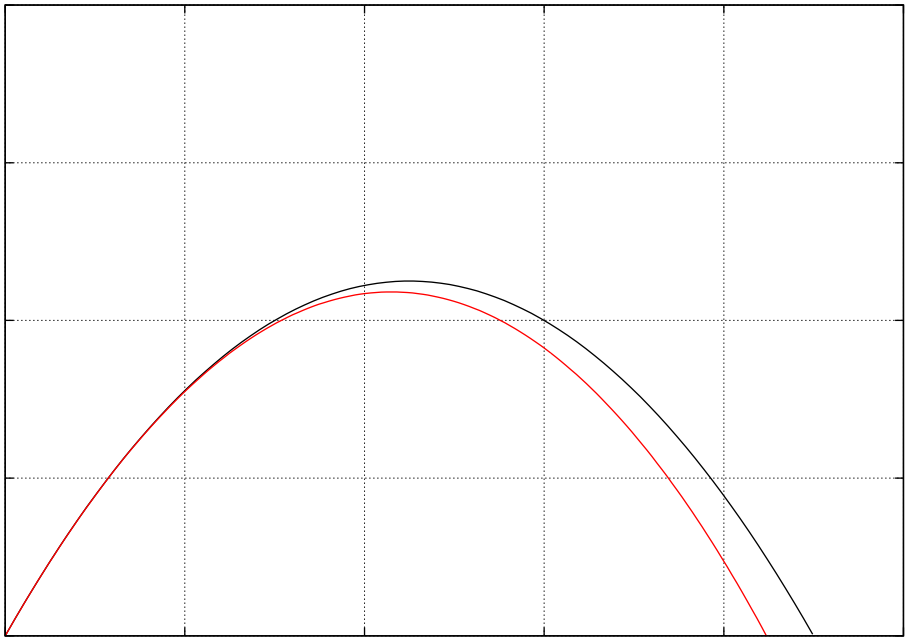


Figure A.2: Air Resistance Proportional to Squared Velocity

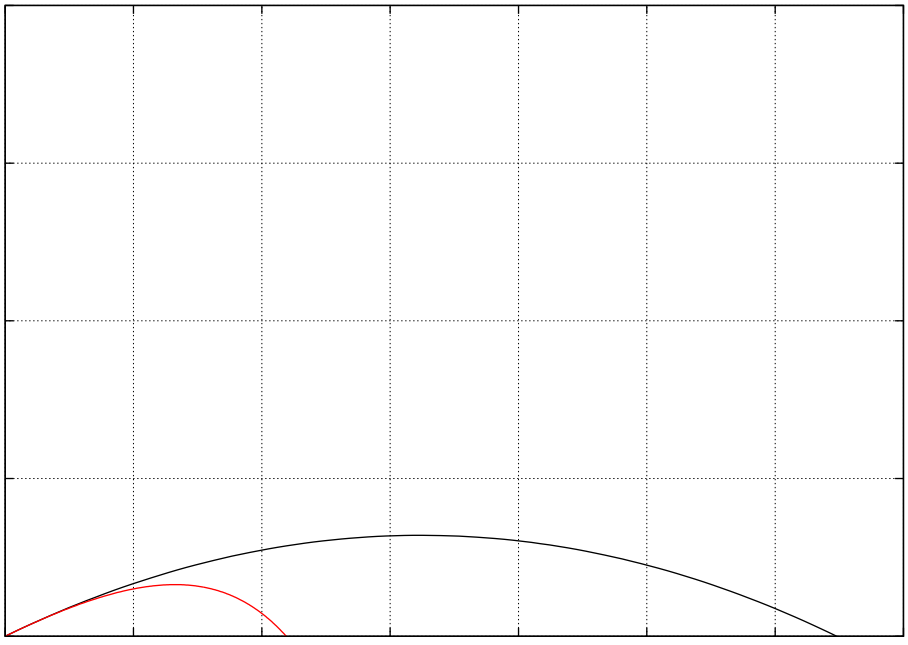


Figure A.3: Trajectory of a 30-06 Round

Appendix B

Energy-Momentum Conservation from Translational Invariance

To follow the derivation in this appendix, you need to be familiar with the Lagrangian dynamics covered in §21.3.

In physical terms, conservation of momentum and energy are a direct consequence of the laws of physics being independent of where you are and what day it is: physics is the same over here as it is over there and the same today as it was yesterday and will be tomorrow. In other words, making a shift (a translation) along any of the three spatial axes or along the time axis should make no difference in the physics you observe. The object of the game is therefore to show that the invariance of physics under spacetime translations leads to a quantity that has a vanishing spacetime derivative and hence is conserved. The derivation of this quantity, known as the energy-momentum tensor, is similar for all the various fields corresponding to matter and forces, though for some kinds of fields there are technical complications; for simplicity we will work through the derivation only for the case of a scalar field ϕ.

Now, our field ϕ will be a function of spacetime location (ct, x, y, z):[1] $\phi = \phi(ct, x, y, z)$. The big-people way to write the four-component vector or four-vector (ct, x, y, z) is x^μ, where the index μ runs from 0 to 3, with x^0 corresponding to ct and x^1, x^2, x^3 to x, y, z. Though it might seem a bit confusing at first, when this four-vector x^μ is used as the argument of a

[1] We include a factor of the speed of light c with the time coordinate t in order to make the dimensions of the coordinates consistent: ct has length dimensions. This is necessary because these coordinates transform into each other as we go from one reference frame to another under the Lorentz transforms of relativity.

function it is customary to omit the index: thus $\phi = \phi(x)$ indicates that ϕ is a function of the spacetime coordinate x^μ. And while we're at it, we might as well also note that with what is known as the summation convention, an index that is repeated in a term is understood to be summed over. Thus, for example, $\partial x^\mu / \partial x^\mu$ stands for

$$\frac{\partial x^\mu}{\partial x^\mu} = \sum_{\mu=0}^{3} \frac{\partial x^\mu}{\partial x^\mu} = \frac{\partial (ct)}{\partial (ct)} + \frac{\partial x}{\partial x} + \frac{\partial y}{\partial y} + \frac{\partial z}{\partial z} = 1 + 1 + 1 + 1 = 4$$

But back to our field $\phi(x)$: if we make an infinitesimal shift

$$x^\mu \to x^\mu + \delta x^\mu$$

in our spacetime location, that is, of we do

$$(ct, x, y, z) \to (ct + c\,\delta t,\ x + \delta x,\ y + \delta y,\ z + \delta z)$$

then the corresponding effect on the field ϕ is to shift it to the value it takes at this shifted spacetime location:

$$\phi(ct, x, y, z) \to \phi(ct + c\,\delta t,\ x + \delta x,\ y + \delta y,\ z + \delta z)$$

Since the shift we are making is infinitesimal, we can neglect the higher powers of δt, δx, δy, and δz in the Taylor expansion of ϕ, so that we have

$$\begin{aligned} \phi(ct, x, y, z) &\to \phi(ct + c\,\delta t,\ x + \delta x,\ y + \delta y,\ z + \delta z) \\ &= \phi(ct, x, y, z) + \frac{\partial \phi}{c\,\partial t}\, c\,\delta t + \frac{\partial \phi}{\partial x}\,\delta x + \frac{\partial \phi}{\partial y}\,\delta y + \frac{\partial \phi}{\partial z}\,\delta z \end{aligned}$$

Using four-vector notation and the summation convention, this may be much more succinctly expressed as

$$\phi(x) \to \phi(x) + \frac{\partial \phi}{\partial x^\mu}\,\delta x^\mu$$

Thus the shift $\delta\phi$ in the value of the field ϕ is

$$\delta\phi = \phi(x + \delta x) - \phi(x) = \frac{\partial \phi}{\partial x^\mu}\,\delta x^\mu \tag{B.1}$$

The Lagrangian [2] $\mathcal{L}$ for our field ϕ is a function of ϕ and its derivatives, which in turn are functions of the spacetime location x^μ, so our infinitesimal shift

[2] Actually, this is the Lagrangian density, which is just the Lagrangian per unit volume. Since a field extends over all of space, to work with the values of the field at specific points in space we work with quantities like energy per unit volume, similar to the way we worked with charge density ρ and current density $\mathbf{j}$ when dealing with electric and magnetic fields in the Maxwell equations. But working per unit volume doesn't change the Lagrange relations. The fancy $\mathcal{L}$ that we have used is the conventional symbol for a Lagrangian density, to distinguish it from an ordinary Lagrangian L.

in spacetime location will have a similar effect on the value of $\mathcal{L}$:

$$\delta\mathcal{L} = \frac{\partial\mathcal{L}}{\partial x^\mu}\,\delta x^\mu \tag{B.2}$$

The Lagrangian does not, however, depend explicitly on the spacetime coordinate x^μ, only implicitly through its dependence on ϕ and the derivatives of ϕ, so that

$$\mathcal{L} = \mathcal{L}\left(\phi, \frac{\partial\phi}{\partial x^\mu}\right)$$

and therefore $\delta\mathcal{L}$ can also be expressed as

$$\delta\mathcal{L} = \delta\mathcal{L}\left(\phi, \frac{\partial\phi}{\partial x^\mu}\right) = \frac{\partial\mathcal{L}}{\partial\phi}\,\delta\phi + \frac{\partial\mathcal{L}}{\partial\frac{\partial\phi}{\partial x^\mu}}\,\delta\frac{\partial\phi}{\partial x^\mu}$$

To keep from getting writer's cramp (or, these days, carpal-tunnel syndrome) it is customary to use a condensed notation for derivatives with respect to spacetime coordinates:[3]

$$\frac{\partial\phi}{\partial x^\mu} = \partial_\mu\phi$$

With this condensed notation, we have

$$\delta\mathcal{L} = \frac{\partial\mathcal{L}}{\partial\phi}\,\delta\phi + \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,\delta\partial_\mu\phi \tag{B.3}$$

Putting eqq. (B.2) and (B.3) together, we arrive at

$$\frac{\partial\mathcal{L}}{\partial x^\mu}\,\delta x^\mu = \frac{\partial\mathcal{L}}{\partial\phi}\,\delta\phi + \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,\delta\partial_\mu\phi$$

or, if we move everything to the right-hand side,

$$0 = \frac{\partial\mathcal{L}}{\partial\phi}\,\delta\phi + \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,\delta\partial_\mu\phi - \frac{\partial\mathcal{L}}{\partial x^\mu}\,\delta x^\mu \tag{B.4}$$

Remember now that our object is to arrive at a quantity with a vanishing spacetime derivative. We therefore want to gerrymander eq. (B.4) until we can pull out an overall δx, in the hope that what multiplies the δx will in fact prove to be a total derivative. To this end, we first use eq. (B.1) to rewrite the $\delta\phi$ and $\delta\partial_\mu\phi$ in terms of δx:

$$\delta\phi = \frac{\partial\phi}{\partial x^\mu}\,\delta x^\mu = \partial_\mu\phi\,\delta x^\mu$$

[3] The reason for the change from superscript to subscript has to do with contravariant and covariant components. This is explained on p.1080 in §22.6.1, but you don't really need to worry about this for our present purposes.

$$\delta\partial_\mu\phi = \delta\frac{\partial\phi}{\partial x^\mu} = \frac{\partial}{\partial x^\nu}\left(\frac{\partial\phi}{\partial x^\mu}\right)\delta x^\nu = \frac{\partial^2\phi}{\partial x^\mu\partial x^\nu}\,\delta x^\nu = \partial_\mu\partial_\nu\phi\,\delta x^\nu$$

Thus eq. (B.4) becomes

$$0 = \frac{\partial\mathcal{L}}{\partial\phi}\,\partial_\mu\phi\,\delta x^\mu + \frac{\partial\mathcal{L}}{\partial(\partial_\mu\phi)}\,\partial_\mu\partial_\nu\phi\,\delta x^\nu - \partial_\mu\mathcal{L}\,\delta x^\mu \tag{B.5}$$

We can rewrite all this in terms of a total derivative if we use the equation of motion for ϕ, which will, similar to the usual

$$\frac{d}{dt}\left(\frac{dL}{d\dot{q}}\right) - \frac{dL}{dq} = 0$$

that we obtain for $L = L(q,\dot{q})$, be

$$\frac{\partial}{\partial x^\nu}\left(\frac{\partial\mathcal{L}}{\partial\left(\frac{\partial\phi}{\partial x^\nu}\right)}\right) - \frac{\partial\mathcal{L}}{\partial\phi} = 0$$

or

$$\partial_\nu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\nu\phi)}\right) - \frac{\partial\mathcal{L}}{\partial\phi} = 0 \tag{B.6}$$

This equation of motion, like all equations of motion, comes from minimizing the action $S = \int dt\,L$, although for fields that are functions of spacetime location the $\mathcal{L}$ that we have been calling the Lagrangian is really the Lagrangian density, as noted in footnote 2. That is, $\mathcal{L}$ is the Lagrangian per unit volume, so that the action for a field is actually the integral over the spatial coordinates as well as time:

$$S = \int c\,dt\,dx\,dy\,dz\;\mathcal{L} = \int d^4x\,\mathcal{L}$$

Hence the ∂_ν in eq. (B.6): instead of just a time derivative, minimizing the action involves derivatives with respect to all of the spacetime coordinates. If we use eq. (B.6), the first term in eq. (B.5) can be re-expressed as

$$\frac{\partial\mathcal{L}}{\partial\phi}\,\partial_\mu\phi\,\delta x^\mu = \left(\partial_\nu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\nu\phi)}\right)\right)\partial_\mu\phi\,\delta x^\mu$$

It will, however, be convenient to swap the μ and ν indices on the right-hand side of this relation, which we can freely do because μ and ν, being repeated, are summed over:

$$\left(\partial_\nu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\nu\phi)}\right)\right)\partial_\mu\phi\,\delta x^\mu = \sum_{\mu=0}^{3}\sum_{\nu=0}^{3}\left(\partial_\nu\left(\frac{\partial\mathcal{L}}{\partial(\partial_\nu\phi)}\right)\right)\partial_\mu\phi\,\delta x^\mu$$

$$= \sum_{\mu=0}^{3} \sum_{\nu=0}^{3} \left(\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right) \right) \partial_\nu \phi \, \delta x^\nu$$
$$= \left(\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right) \right) \partial_\nu \phi \, \delta x^\nu$$

Using this in eq. (B.5), as well as

$$A \, \partial_\mu B + B \, \partial_\mu A = \partial_\mu (AB)$$

with

$$A = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \qquad\qquad B = \partial_\nu \phi$$

we obtain

$$0 = \left(\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right) \right) \partial_\nu \phi \, \delta x^\nu + \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \, \partial_\mu \partial_\nu \phi \, \delta x^\nu - \partial_\mu \mathcal{L} \, \delta x^\mu$$
$$= \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \, \partial_\nu \phi \right) \delta x^\nu - \partial_\mu \mathcal{L} \, \delta x^\mu \tag{B.7}$$

We are almost there: all that prevents us from pulling out an overall factor of δx from a total derivative with respect to x^μ (the ∂_μ) is that the first term involves a δx^ν while the second involves δx^μ. To get around this, we use the metric $\eta_{\mu\nu}$ discussed on p.1079f. in §22.6.1: [4]

$$\delta x^\mu = \eta^\mu{}_\nu \, \delta x^\nu$$

Thus eq. (B.7) yields

$$0 = \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \, \partial_\nu \phi \right) \delta x^\nu - \partial_\mu \mathcal{L} \, \eta^\mu{}_\nu \, \delta x^\nu$$
$$= \partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \, \partial_\nu \phi - \eta^\mu{}_\nu \, \mathcal{L} \right) \delta x^\nu$$

which, since this must hold for all δx^ν, means that

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \, \partial_\nu \phi - \eta^\mu{}_\nu \, \mathcal{L} \right) = 0$$

[4] It may seem like a real rip-off to have followed along this far and then, at the very end, to have this hand-waving about the metric, but we are only playing games to get the indices to cooperate, and you should not allow this to obscure the main point of the derivation: that the invariance of physics under spacetime translations leads, in the Lagrangian formalism, to a quantity that is conserved in the sense that its spacetime derivative vanishes.

This leads us to define the energy-momentum tensor for the field ϕ to be

$$T^{\mu}{}_{\nu} = \frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \, \partial_\nu \phi - \eta^{\mu}{}_{\nu} \, \mathcal{L}$$

which satisfies the conservation relation

$$\partial_\mu T^{\mu}{}_{\nu} = 0$$

The mathematician Emma Noether proved that for every symmetry in the Lagrangian, there is a corresponding conserved "charge" and "current." In the case of invariance under spacetime translations, that conserved quantity is energy and momentum.

Appendix C

Proofs of Kepler's Laws

C.1 The First Law

In this section we prove Kepler's first law: that gravitational orbits are elliptical.

For a satellite of mass m orbiting a large body of mass $M \gg m$, we can to a good approximation treat M as stationary.[1] For simplicity, we can also, without loss of generality, put M at the origin.

The first thing to note is that the energy E and angular momentum L are constants of the motion and will therefore be useful parameters in terms of which to set up our calculation. The energy is constant because the gravitational force is conservative; to see that the angular momentum is also constant, recall that it is torque, in this case the torque due to the gravitational force, that determines the rate of change of angular momentum:

$$\frac{d\mathbf{L}}{dt} = \boldsymbol{\tau} = \mathbf{r} \times \mathbf{F}$$

The gravitational force is, however, a central force: with M at the origin, the gravitational force exerted on m is in the $-\hat{\mathbf{r}}$ direction, so that

$$\mathbf{F} = -\frac{GMm}{r^2}\,\hat{\mathbf{r}}$$

Since $\mathbf{r} \times \hat{\mathbf{r}} = r\,\hat{\mathbf{r}} \times \hat{\mathbf{r}} = 0$, there is no torque due to the gravitational force and angular momentum is conserved.

Polar coordinates will turn out to be most convenient for the calculation. With dots denoting time derivatives, the relations for the angular momentum L and energy E of the satellite are (see eq. (3.41) on p.151)

$$L = I\omega = mr^2\dot{\theta} \tag{C.1a}$$

[1] The proof we give is still valid even without this approximation, as long as you use the reduced mass $\mu = Mm/(M+m)$ in place of m in the nongravitational terms (that is, everywhere except when it occurs in the combination GMm).

$$\begin{aligned} E &= \tfrac{1}{2}mv^2 - \frac{GMm}{r} \\ &= \tfrac{1}{2}m(v_r^2 + v_\theta^2) - \frac{GMm}{r} \\ &= \tfrac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2) - \frac{GMm}{r} \end{aligned} \tag{C.1b}$$

In this form, these equations of motion are relatively intractable: we have coupled differential equations mixing r, $\dot{r}$, and $\dot{\theta}$. But we can separate the r's and θ's to some extent by solving eq. (C.1a) for $\dot{\theta}$, to obtain

$$\dot{\theta} = \frac{L}{mr^2} \tag{C.2a}$$

and then using this result in eq. (C.1b):

$$\begin{aligned} E &= \tfrac{1}{2}m\left(\dot{r}^2 + r^2\left(\frac{L}{mr^2}\right)^2\right) - \frac{GMm}{r} \\ &= \tfrac{1}{2}m\dot{r}^2 + \frac{L^2}{2mr^2} - \frac{GMm}{r} \end{aligned} \tag{C.2b}$$

At least this latter equation now involves only r and $\dot{r}$. If we solve eq. (C.2b) for $\dot{r}$, our equations of motion become

$$\dot{\theta} = \frac{L}{mr^2} \tag{C.3a}$$

$$\dot{r} = \sqrt{\frac{2}{m}\left(E - \frac{L^2}{2mr^2} + \frac{GMm}{r}\right)} \tag{C.3b}$$

In principle we could solve eq. (C.3b) to get $r(t)$ and, using this solution for r in eq. (C.3a), then solve for $\theta(t)$, but we want rather to determine the *shape* of the orbit, that is, to obtain a result for r not as a function of time but as a function of θ: $r = r(\theta)$. The easy way to eliminate time from eqq. (C.3) and get a relation between just r and θ is to divide the $\dot{r}$ of eq. (C.3b) by the $\dot{\theta}$ of eq. (C.3a): the dt's cancel, leaving us with

$$\frac{dr}{d\theta} = \frac{mr^2}{L}\sqrt{\frac{2}{m}\left(E - \frac{L^2}{2mr^2} + \frac{GMm}{r}\right)}$$

To integrate this, we simply separate variables in the usual way:

$$\int_{\theta_0}^{\theta} d\theta = \int_{r_0}^{r} dr\ \frac{L}{mr^2}\left[\frac{2}{m}\left(E - \frac{L^2}{2mr^2} + \frac{GMm}{r}\right)\right]^{-\frac{1}{2}}$$

The $d\theta$ integration is mercifully easy. The dr integration is a different story. But if you stare at the dr integration hard enough, you can see that the dr/r^2 suggests the substitution

$$u = \frac{1}{r} \qquad\qquad du = -\frac{dr}{r^2} \tag{C.4}$$

With this substitution, the above integrations become

$$\begin{aligned}
\theta - \theta_0 &= -\int_{u_0}^{u} du\, \frac{L}{m}\left[\frac{2}{m}\left(E - \frac{L^2}{2m}u^2 + GMmu\right)\right]^{-\frac{1}{2}} \\
&= -\int_{u_0}^{u} du\, \left[\frac{m^2}{L^2}\frac{2}{m}\left(E - \frac{L^2}{2m}u^2 + GMmu\right)\right]^{-\frac{1}{2}} \\
&= -\int_{u_0}^{u} du\, \left[\frac{2mE}{L^2} - u^2 + \frac{2GMm^2}{L^2}u\right]^{-\frac{1}{2}}
\end{aligned}$$

If we complete the square, this "simplifies" to

$$\theta - \theta_0 = -\int_{u_0}^{u} du\, \left[\left(\frac{2mE}{L^2} + \left(\frac{GMm^2}{L^2}\right)^2\right) - \left(u - \frac{GMm^2}{L^2}\right)^2\right]^{-\frac{1}{2}}$$

and we are now in a position to do the integration by trig substitution: if we define new animals z and β by

$$\beta = \sqrt{\frac{2mE}{L^2} + \left(\frac{GMm^2}{L^2}\right)^2} \qquad z = u - \frac{GMm^2}{L^2} \qquad dz = du \tag{C.5}$$

then we have

$$\theta - \theta_0 = -\int_{z_0}^{z} dz\, \frac{1}{\sqrt{\beta^2 - z^2}}$$

which, after the trig substitution

$$z = \beta\cos\psi \qquad\qquad dz = -\beta\sin\psi\, d\psi \tag{C.6}$$

becomes

$$\begin{aligned}
\theta - \theta_0 &= -\int_{\psi_0}^{\psi} -\beta\sin\psi\, d\psi\, \frac{1}{\sqrt{\beta^2 - \beta^2\cos^2\psi}} \\
&= \int_{\psi_0}^{\psi} d\psi\, \frac{\sin\psi}{\sqrt{1 - \cos^2\psi}} \\
&= \int_{\psi_0}^{\psi} d\psi\, \frac{\sin\psi}{\sin\psi} \\
&= \psi - \psi_0
\end{aligned}$$

Now that the integrations are done, we just have to reverse the chain of substitutions (C.6), (C.5), and (C.4):

$$\begin{aligned}\theta - \theta_0 &= \psi - \psi_0 \\ &= \cos^{-1}\left(\frac{z}{\beta}\right) - \cos^{-1}\left(\frac{z_0}{\beta}\right) \\ &= \cos^{-1}\left(\frac{u - GMm^2/L^2}{\beta}\right) - \cos^{-1}\left(\frac{u_0 - GMm^2/L^2}{\beta}\right) \\ &= \cos^{-1}\left(\frac{1/r - GMm^2/L^2}{\beta}\right) - \cos^{-1}\left(\frac{1/r_0 - GMm^2/L^2}{\beta}\right) \qquad \text{(C.7)}\end{aligned}$$

where we have for the time being left things in terms of the constant β of eqq. (C.5) just to keep the expressions inside the inverse cosines from becoming algebraic monstrosities.

Our result (C.7) probably doesn't look very elliptical to you, so we still have some work ahead of us. First of all, in the lower limits of our integrations θ_0 was the angle corresponding to r_0, but since we can measure our angles from any $\theta = 0$ that we want, we can simplify (C.7) considerably by setting

$$\theta_0 = \cos^{-1}\left(\frac{1/r_0 - GMm^2/L^2}{\beta}\right)$$

With this choice, the second terms on the right and left sides of eq. (C.7) will cancel each other and we will be left with just

$$\theta = \cos^{-1}\left(\frac{1/r - GMm^2/L^2}{\beta}\right)$$

When solved for r, this yields

$$r = \frac{1}{\dfrac{GMm^2}{L^2} + \beta\cos\theta}$$

or, if we now put in the value of β from eqq. (C.5),

$$r = \frac{1}{\dfrac{GMm^2}{L^2} + \sqrt{\dfrac{2mE}{L^2} + \left(\dfrac{GMm^2}{L^2}\right)^2}\,\cos\theta}$$

Multiplying both numerator and denominator by L^2/GMm^2, we can simplify this to

$$r = \frac{\dfrac{L^2}{GMm^2}}{1 + \sqrt{\left(\dfrac{L^2}{GMm^2}\right)^2\left(\dfrac{2mE}{L^2} + \left(\dfrac{GMm^2}{L^2}\right)^2\right)}\,\cos\theta}$$

$$= \frac{L^2/GMm^2}{1 + \sqrt{1 + \dfrac{2EL^2}{m(GMm)^2}}\ \cos\theta} \tag{C.8}$$

This relation for r corresponds to an ellipse of eccentricity

$$\epsilon = \sqrt{1 + \frac{2EL^2}{m(GMm)^2}} \tag{C.9}$$

And there you have it. Maybe. Or maybe eq. (C.8) still doesn't look very elliptical to you. You may be more familiar with

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1 \tag{C.10}$$

as the relation for an ellipse. We can, however, gerrymander eq. (C.8) into eq. (C.10) without too much difficulty. First, for simplicity let us write eq. (C.8) in the form

$$r = \frac{\rho}{1 + \epsilon\cos\theta} \tag{C.11}$$

where the eccentricity ϵ is given by eq. (C.9) and where we have defined

$$\rho = \frac{L^2}{GMm^2} \tag{C.12}$$

If now we multiply both sides of eq. (C.11) by $1 + \epsilon\cos\theta$ and use $r\cos\theta = x$, we obtain

$$\begin{aligned} r + \epsilon r\cos\theta &= \rho \\ r + \epsilon x &= \rho \\ r &= \rho - \epsilon x \end{aligned}$$

Squaring both sides and using $r^2 = x^2 + y^2$, we further have

$$\begin{aligned} r^2 &= (\rho - \epsilon x)^2 \\ x^2 + y^2 &= \rho^2 - 2\epsilon\rho x + \epsilon^2 x^2 \end{aligned}$$

Next, we gather terms in x and y and complete the square of the terms in x by adding

$$\frac{\epsilon^2\rho^2}{1 - \epsilon^2}$$

to both sides:

$$(1 - \epsilon^2)x^2 + 2\epsilon\rho x + y^2 = \rho^2$$

$$(1-\epsilon^2)\left(x^2+\frac{2\epsilon\rho}{1-\epsilon^2}x+\frac{\epsilon^2\rho^2}{(1-\epsilon^2)^2}\right)+y^2=\rho^2+\frac{\epsilon^2\rho^2}{1-\epsilon^2}$$

$$(1-\epsilon^2)\left(x+\frac{\epsilon\rho}{1-\epsilon^2}\right)^2+y^2=\frac{\rho^2}{1-\epsilon^2} \tag{C.13}$$

This is getting close to eq. (C.10), but is in terms of a shifted x coordinate. This is of course because the ellipse of eq. (C.10) is centered at the origin, while that of eq. (C.8) is not: with eq. (C.8), we have the biased values

$$r\Big|_{\theta=0}=\frac{\rho}{1+\epsilon} \qquad\qquad r\Big|_{\theta=\pi}=\frac{\rho}{1-\epsilon}$$

If we define a shifted coordinate

$$x'=x+\frac{\epsilon\rho}{1-\epsilon^2}$$

then eq. (C.13) simplifies to

$$(1-\epsilon^2)x'^2+y^2=\frac{\rho^2}{1-\epsilon^2}$$

or, if we divide both sides by the right-hand side,

$$\left(\frac{x'}{\rho/(1-\epsilon^2)}\right)^2+\left(\frac{y}{\rho/\sqrt{1-\epsilon^2}}\right)^2=1$$

which is indeed in the form (C.10), with

$$a=\frac{\rho}{1-\epsilon^2} \qquad\qquad b=\frac{\rho}{\sqrt{1-\epsilon^2}} \tag{C.14}$$

C.2 The Second Law

You will doubtless be relieved to hear that the proofs of the second and third laws are much easier.

Kepler's second law states that equal areas are swept out in equal times. We therefore want to derive an expression for the areal velocity, that is, for the area swept out per unit time. During an infinitesimal time interval dt, the orbiting mass m will sweep out an infinitesimal "pizza slice" of radius r and arc length $r\,d\theta$. If we treat this slice as a triangle of base $r\,d\theta$ and altitude r, the area swept out will be

$$dA=\tfrac{1}{2}r(r\,d\theta)=\tfrac{1}{2}r^2\,d\theta$$

The slice is of course not triangular, nor even, because r is not constant, circular, but the corrections for these deviations from triangular shape would

be infinitesimal in comparison to our $dA = \frac{1}{2}r^2\,d\theta$ and may therefore be neglected. The rate at which area is swept out is therefore

$$\frac{dA}{dt} = \tfrac{1}{2}r^2\,\frac{d\theta}{dt}$$

Recall now that, according to eq. (C.3a),

$$\frac{d\theta}{dt} = \frac{L}{mr^2}$$

We can therefore re-express the areal velocity as

$$\frac{dA}{dt} = \tfrac{1}{2}r^2\frac{L}{mr^2} = \frac{L}{2m} \tag{C.15}$$

Since the angular momentum L is constant, this means that area is swept out at a constant rate.

C.3 The Third Law

To prove Kepler's third law, that the square of the period T is proportional to the cube of the semimajor axis a, we first relate the period to the axes of the ellipse by integrating our result (C.15) for the areal velocity and recalling that the area of an ellipse is πab: since both L and m are constant in

$$dA = \frac{L}{2m}\,dt$$

integrating over one full cycle yields

$$\int dA = \int \frac{L}{2m}\,dt$$
$$\pi ab = \frac{L}{2m}\,T$$

When solved for T^2, this gives

$$T^2 = \left(\frac{2\pi mab}{L}\right)^2 \tag{C.16}$$

We now want to gerrymander the right-hand side of eq. (C.16) into the cube of the semimajor axis a and, we hope, a factor that involves only quantities that are the same for the orbits of all the planets. And eqq. (C.14) and (C.12) allow us to accomplish this relatively painlessly:

$$T^2 = \left(\frac{2\pi m}{L}\,ab\right)^2$$

$$
\begin{aligned}
&= \left(\frac{2\pi m}{L} \frac{\rho}{1-\epsilon^2} \frac{\rho}{\sqrt{1-\epsilon^2}} \right)^2 \\
&= \frac{4\pi^2 m^2}{L^2} \frac{\rho^4}{(1-\epsilon^2)^3} \\
&= \frac{4\pi^2 m^2}{L^2} \rho \left(\frac{\rho}{1-\epsilon^2} \right)^3 \\
&= \frac{4\pi^2 m^2}{L^2} \rho\, a^3 \\
&= \frac{4\pi^2 m^2}{L^2} \frac{L^2}{GMm^2} a^3 \\
&= \frac{4\pi^2}{GM} a^3
\end{aligned}
$$

Bingo: G and M are indeed the same for the orbits of all the planets (as opposed to quantities like m and L, which vary from one planet's orbit to another's). We have thus established Kepler's third law.

Appendix D

Proof of the Linearity of Lorentz Transforms

To prove that Lorentz transforms are linear without going through ridiculously extreme gymnastics, we will need to use what is known as four-vector notation: the big-people way to denote the four-dimensional spacetime vector (ct, x, y, z) is x^μ, where the index μ runs from 0 to 3, with x^0 corresponding to ct and x^1, x^2, x^3 to x, y, z. We will also use what is known as the summation convention: in terms where an index is repeated, that index is understood to be summed over. Thus, for example, $\partial x^\mu / \partial x^\mu$ stands for

$$\begin{aligned}\frac{\partial x^\mu}{\partial x^\mu} &= \sum_{\mu=0}^{3} \partial_\mu x^\mu \\ &= \frac{\partial (ct)}{\partial (ct)} + \frac{\partial x}{\partial x} + \frac{\partial y}{\partial y} + \frac{\partial z}{\partial z} \\ &= 1 + 1 + 1 + 1 \\ &= 4\end{aligned}$$

In special relativity, the square $c^2\, d\tau^2$ of the invariant interval,

$$\begin{aligned}c^2\, d\tau^2 &= (c\, dt)^2 - dx^2 - dy^2 - dz^2 \\ &= (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2\end{aligned}$$

can be neatly written as

$$c^2\, d\tau^2 = \eta_{\mu\nu}\, dx^\mu\, dx^\nu \tag{D.1}$$

where $\eta_{\mu\nu}$ is the matrix, known as the flat-space metric, defined as

$$\eta_{\mu\nu} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}$$

To see this, we have only to do out the multiplication in eq. (D.1):

$$\begin{aligned}
\eta_{\mu\nu}\,dx^\mu\,dx^\nu &= dx^\mu\,\eta_{\mu\nu}\,dx^\nu \\
&= \sum_{\mu=0}^{3}\sum_{\nu=0}^{3} dx^\mu\,\eta_{\mu\nu}\,dx^\nu \\
&= \begin{pmatrix} dx^0 & dx^1 & dx^2 & dx^3 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} dx^0 \\ dx^1 \\ dx^2 \\ dx^3 \end{pmatrix} \\
&= \begin{pmatrix} dx^0 & dx^1 & dx^2 & dx^3 \end{pmatrix} \begin{pmatrix} dx^0 & 0 & 0 & 0 \\ 0 & -dx^1 & 0 & 0 \\ 0 & 0 & -dx^2 & 0 \\ 0 & 0 & 0 & -dx^3 \end{pmatrix} \\
&= (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2 \quad \checkmark
\end{aligned}$$

If x^μ and x'^μ are the coordinates in the reference frames F and F', respectively, then, since $d\tau^2$ is invariant, we must have

$$\eta_{\mu\nu}\,dx^\mu\,dx^\nu = \eta_{\mu\nu}\,dx'^\mu\,dx'^\nu \tag{D.2}$$

But by some basic calculus, we must also have

$$\begin{aligned}
dx'^\mu &= \frac{\partial x'^\mu}{\partial x^0}\,dx^0 + \frac{\partial x'^\mu}{\partial x^1}\,dx^1 + \frac{\partial x'^\mu}{\partial x^2}\,dx^2 + \frac{\partial x'^\mu}{\partial x^3}\,dx^3 \\
&= \sum_{\lambda=0}^{3} \frac{\partial x'^\mu}{\partial x^\lambda}\,dx^\lambda \\
&= \frac{\partial x'^\mu}{\partial x^\lambda}\,dx^\lambda
\end{aligned}$$

where in the last step we have used the summation convention. And likewise we will also have

$$dx'^\nu = \frac{\partial x'^\nu}{\partial x^\kappa}\,dx^\kappa$$

Thus we can rewrite eq. (D.2) as

$$\begin{aligned}
\eta_{\mu\nu}\,dx^\mu\,dx^\nu &= \eta_{\mu\nu}\frac{\partial x'^\mu}{\partial x^\lambda}\,dx^\lambda\frac{\partial x'^\nu}{\partial x^\kappa}\,dx^\kappa \\
&= \eta_{\mu\nu}\frac{\partial x'^\mu}{\partial x^\lambda}\frac{\partial x'^\nu}{\partial x^\kappa}\,dx^\lambda\,dx^\kappa
\end{aligned} \tag{D.3}$$

Now, the partial derivatives $\partial x'^\mu/\partial x^\lambda$ that occur in relations like

$$dx'^\mu = \frac{\partial x'^\mu}{\partial x^\lambda}\,dx^\lambda$$

take us from the coordinate differentials dx to the coordinate differentials dx'. In other words, these derivatives *are* the transform between the coordinates of F and the coordinates of F'. We want to prove that this transform is linear. To accomplish this, we will show that the derivatives of this transform vanish, that is, that

$$\frac{\partial}{\partial x^\rho}\frac{\partial x'^\mu}{\partial x^\lambda} = 0 \tag{D.4}$$

First, note that it doesn't matter how an index that is summed over is labeled, in the sense that, for example,

$$\sum_{n=0}^{3} x_n^2 = \sum_{m=0}^{3} x_m^2$$

Such repeated indices are called dummy indices, and they may therefore be relabeled with the same freedom that dummy variables can be relabeled in integrations. So there is no reason why we can't relabel the μ and ν on the left-hand side of eq. (D.3) as λ and κ:

$$\eta_{\lambda\kappa}\, dx^\lambda\, dx^\kappa = \eta_{\mu\nu}\frac{\partial x'^\mu}{\partial x^\lambda}\frac{\partial x'^\nu}{\partial x^\kappa}\, dx^\lambda\, dx^\kappa$$

Since this relation must hold for all dx^λ and dx^κ, we must have

$$\eta_{\lambda\kappa} = \eta_{\mu\nu}\frac{\partial x'^\mu}{\partial x^\lambda}\frac{\partial x'^\nu}{\partial x^\kappa} \tag{D.5}$$

To generate derivatives like those in eq. (D.4), and hopefully some relation that will enable to show that these derivatives vanish, we can apply $\partial/\partial x^\rho$ to both sides of eq. (D.5):

$$\begin{aligned}\frac{\partial \eta_{\lambda\kappa}}{\partial x^\rho} &= \frac{\partial}{\partial x^\rho}\left(\eta_{\mu\nu}\frac{\partial x'^\mu}{\partial x^\lambda}\frac{\partial x'^\nu}{\partial x^\kappa}\right)\\ &= \frac{\partial \eta_{\mu\nu}}{\partial x^\rho}\frac{\partial x'^\mu}{\partial x^\lambda}\frac{\partial x'^\nu}{\partial x^\kappa} + \eta_{\mu\nu}\frac{\partial^2 x'^\mu}{\partial x^\rho \partial x^\lambda}\frac{\partial x'^\nu}{\partial x^\kappa} + \eta_{\mu\nu}\frac{\partial x'^\mu}{\partial x^\lambda}\frac{\partial^2 x'^\nu}{\partial x^\rho \partial x^\kappa}\end{aligned}$$

This may look like an intractable mess, but since η is a constant matrix — all of its elements are constants — the derivatives of η vanish, leaving us with

$$0 = \eta_{\mu\nu}\frac{\partial^2 x'^\mu}{\partial x^\rho \partial x^\lambda}\frac{\partial x'^\nu}{\partial x^\kappa} + \eta_{\mu\nu}\frac{\partial x'^\mu}{\partial x^\lambda}\frac{\partial^2 x'^\nu}{\partial x^\rho \partial x^\kappa} \tag{D.6a}$$

This may still look like an intractable mess, but now comes the magic: if we interchange the indices λ and ρ — there was nothing special about our choice

of those labels to begin with — , we can equivalently write this relation as[1]

$$0 = \eta_{\mu\nu}\frac{\partial^2 x'^{\mu}}{\partial x^{\lambda}\partial x^{\rho}}\frac{\partial x'^{\nu}}{\partial x^{\kappa}} + \eta_{\mu\nu}\frac{\partial x'^{\mu}}{\partial x^{\rho}}\frac{\partial^2 x'^{\nu}}{\partial x^{\lambda}\partial x^{\kappa}} \tag{D.6b}$$

Similarly, if we interchange the indices ρ and κ, we can also write eq. (D.6a) as

$$0 = \eta_{\mu\nu}\frac{\partial^2 x'^{\mu}}{\partial x^{\kappa}\partial x^{\lambda}}\frac{\partial x'^{\nu}}{\partial x^{\rho}} + \eta_{\mu\nu}\frac{\partial x'^{\mu}}{\partial x^{\lambda}}\frac{\partial^2 x'^{\nu}}{\partial x^{\kappa}\partial x^{\rho}} \tag{D.6c}$$

If, with almost divine foresight, we add eq. (D.6b) to (D.6a) and subtract eq. (D.6c), we obtain

$$\begin{aligned} 0 = {} & \eta_{\mu\nu}\frac{\partial^2 x'^{\mu}}{\partial x^{\rho}\partial x^{\lambda}}\frac{\partial x'^{\nu}}{\partial x^{\kappa}} + \eta_{\mu\nu}\frac{\partial x'^{\mu}}{\partial x^{\lambda}}\frac{\partial^2 x'^{\nu}}{\partial x^{\rho}\partial x^{\kappa}} \\ & + \eta_{\mu\nu}\frac{\partial^2 x'^{\mu}}{\partial x^{\lambda}\partial x^{\rho}}\frac{\partial x'^{\nu}}{\partial x^{\kappa}} + \eta_{\mu\nu}\frac{\partial x'^{\mu}}{\partial x^{\rho}}\frac{\partial^2 x'^{\nu}}{\partial x^{\lambda}\partial x^{\kappa}} \\ & - \eta_{\mu\nu}\frac{\partial^2 x'^{\mu}}{\partial x^{\kappa}\partial x^{\lambda}}\frac{\partial x'^{\nu}}{\partial x^{\rho}} - \eta_{\mu\nu}\frac{\partial x'^{\mu}}{\partial x^{\lambda}}\frac{\partial^2 x'^{\nu}}{\partial x^{\kappa}\partial x^{\rho}} \end{aligned} \tag{D.7}$$

If you look closely, you will see that the second and the last terms cancel out and that the first and the third terms are the same. Thus we have

$$0 = 2\eta_{\mu\nu}\frac{\partial^2 x'^{\mu}}{\partial x^{\rho}\partial x^{\lambda}}\frac{\partial x'^{\nu}}{\partial x^{\kappa}} + \eta_{\mu\nu}\frac{\partial x'^{\mu}}{\partial x^{\rho}}\frac{\partial^2 x'^{\nu}}{\partial x^{\lambda}\partial x^{\kappa}} - \eta_{\mu\nu}\frac{\partial^2 x'^{\mu}}{\partial x^{\kappa}\partial x^{\lambda}}\frac{\partial x'^{\nu}}{\partial x^{\rho}}$$

It takes a little more work, but we can also get the last two of these three remaining terms to cancel out: the indices μ and ν are dummy indices, and we can, if we want, relabel them in the last term by interchanging them. This yields

$$0 = 2\eta_{\mu\nu}\frac{\partial^2 x'^{\mu}}{\partial x^{\rho}\partial x^{\lambda}}\frac{\partial x'^{\nu}}{\partial x^{\kappa}} + \eta_{\mu\nu}\frac{\partial x'^{\mu}}{\partial x^{\rho}}\frac{\partial^2 x'^{\nu}}{\partial x^{\lambda}\partial x^{\kappa}} - \eta_{\nu\mu}\frac{\partial^2 x'^{\nu}}{\partial x^{\kappa}\partial x^{\lambda}}\frac{\partial x'^{\mu}}{\partial x^{\rho}}$$

Since the matrix η is diagonal, $\eta_{\nu\mu} = \eta_{\mu\nu}$, so that we have

$$0 = 2\eta_{\mu\nu}\frac{\partial^2 x'^{\mu}}{\partial x^{\rho}\partial x^{\lambda}}\frac{\partial x'^{\nu}}{\partial x^{\kappa}} + \eta_{\mu\nu}\frac{\partial x'^{\mu}}{\partial x^{\rho}}\frac{\partial^2 x'^{\nu}}{\partial x^{\lambda}\partial x^{\kappa}} - \eta_{\mu\nu}\frac{\partial^2 x'^{\nu}}{\partial x^{\kappa}\partial x^{\lambda}}\frac{\partial x'^{\mu}}{\partial x^{\rho}}$$

Now the last two terms do indeed cancel, and we are left with

$$0 = 2\eta_{\mu\nu}\frac{\partial^2 x'^{\mu}}{\partial x^{\rho}\partial x^{\lambda}}\frac{\partial x'^{\nu}}{\partial x^{\kappa}} \tag{D.8}$$

[1] If this reasoning seems a bit strange to you, think about it this way: on the right-hand side of eq. (D.6a) the indices μ and ν are summed over, so that only κ, λ, and ρ remain. In other words, the right-hand side of eq. (D.6a) is an animal with three indices κ, λ, and ρ; if we call this animal T, we can write eq. (D.6a) as $0 = T_{\kappa\lambda\rho}$. This means that $T_{\kappa\lambda\rho}$ must vanish for all possible sets of values of $\{\kappa, \lambda, \rho\}$ as each of these three indices runs from 0 to 3. But this is the same statement that would be made by the relation $0 = T_{\rho\lambda\kappa}$ or the relation $0 = T_{\kappa\rho\lambda}$. We can then generate eq. (D.7) by adding $0 = T_{\rho\lambda\kappa}$ to $0 = T_{\kappa\lambda\rho}$ and then subtracting $0 = T_{\kappa\rho\lambda}$: $0 = T_{\kappa\lambda\rho} + T_{\rho\lambda\kappa} - T_{\kappa\rho\lambda}$.

In a simple algebraic equation like $abc = 0$, we can reason that if $a \neq 0$ and $c \neq 0$, then we must have $b = 0$. But while $\eta_{\mu\nu} \neq 0$ and $\partial x'^{\nu}/\partial x^{\kappa} \neq 0$ — since $\partial x'^{\nu}/\partial x^{\kappa}$ was our transform between F and F', it wouldn't make any sense for it to vanish — , we have to be a bit more careful here: the right-hand side of eq. (D.8) is not a single term but rather a sum of $4 \times 4 = 16$ terms as we sum over each of the repeated indices μ and ν from 0 to 3. The two "factors," $\eta_{\mu\nu}$ and $\partial x'^{\nu}/\partial x^{\kappa}$, by which

$$\frac{\partial^2 x'^{\mu}}{\partial x^{\rho} \partial x^{\lambda}}$$

is multiplied on the right-hand side of eq. (D.8) are, however, nonsingular matrices, so that they have inverses by which we may multiply both sides of eq. (D.8) — the matrix equivalent of dividing a and c out of $abc = 0$ when $a \neq 0$ and $c \neq 0$. We therefore do indeed have

$$\frac{\partial^2 x'^{\mu}}{\partial x^{\rho} \partial x^{\lambda}} = \frac{\partial}{\partial x^{\rho}} \frac{\partial x'^{\mu}}{\partial x^{\lambda}} = 0 \tag{D.9}$$

This means that the transform $\partial x'^{\mu}/\partial x^{\lambda}$ that leaves the interval invariant is constant. In other words, $\partial x'^{\mu}/\partial x^{\lambda}$ must be a linear in x^{λ}, of the form [2]

$$x'^{\mu} = \Lambda^{\mu}{}_{\lambda} x^{\lambda} + a^{\mu}$$

where $\Lambda^{\mu}{}_{\lambda}$ and a^{μ} are constants. The $\Lambda^{\mu}{}_{\lambda}$ part of this transform is the Lorentz transform that we all know and love. The a^{μ} correspond to constant shifts (translations) in the t, x, y, or z directions — equivalent to shifting the origins of these axes in the frames F and F' relative to each other. The combination of Lorentz invariance and translation invariance is known as Poincaré invariance.

[2] The indices on Λ are written as $\Lambda^{\mu}{}_{\lambda}$, one upper and one lower, for technical reasons that do not concern us here and that therefore will be left mysterious. Those with a low tolerance for the mysterious will find a brief explanation of superscripts and subscripts in terms of contravariant and covariant components on p.1080 in §22.6.1.

Note also that the constant shifts a^{μ} correspond to $\partial x'^{\nu}/\partial x^{\kappa} = 0$ in eq. (D.8).

Appendix E

Proof That
$1+2+3+4+\cdots=-\frac{1}{12}$

The proof that [1]

$$\sum_{n=1}^{\infty} n = 1+2+3+4+\cdots = -\tfrac{1}{12} \tag{E.1}$$

is accomplished by looking at the Riemann zeta function $\zeta(s)$, which is defined by

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} \tag{E.2}$$

In the complex s plane this series expansion is valid, in the sense that it converges, for $\Re(s) > 1$. The sum in eq. (E.1) is $\zeta(-1)$. To get a result for $\zeta(-1)$, we need to find an analytic continuation of $\zeta(s)$: we need to find some other expression for $\zeta(s)$ that is valid when $s = -1$.

Consider, as a simple preliminary example, the expansion (obtainable as a Taylor series or by plain old long division)

$$\frac{1}{1-z} = 1 + z + z^2 + z^3 + \cdots = \sum_{n=0}^{\infty} z^n \tag{E.3}$$

The left-hand side of eq. (E.3) is good as long as $z \neq -1$. The restrictions on the right-hand side are, however, more severe: in the complex plane, the series converges only for points within the unit circle ($|z| < 1$). Although the left- and right-hand sides of eq. (E.3) have different domains, where their domains overlap the two sides of eq. (E.3) represent the same function, and

[1] Some physicists like to facetiously argue that since you are adding all positive numbers, the result must be negative; since you are adding all integers, the result must be a fraction; and that for dimensional reasons the only value this fraction can have is $-\frac{1}{12}$. If you're a physicist, you find this hysterically funny. But then nobody ever said physicists were normal people.

in a rigorous mathematical as well as casual intuitive sense this identity can be extended into regions where only one side of eq. (E.3) is valid. Thus for $|z| \geq 1$ (excluding $z = 1$), the series expansion doesn't converge, but $1/(1-z)$ is perfectly well behaved, and so in this region we take $1/(1-z)$ to represent the function that the series expansion would have represented were it valid. For the region $|z| \geq 1$, $z \neq 1$, $1/(1-z)$ is said to be the analytic continuation of the series on the right-hand side of eq. (E.3). This leads not only to perfectly plausible results like

$$1 + \tfrac{1}{2} + \left(\tfrac{1}{2}\right)^2 + \left(\tfrac{1}{2}\right)^3 + \cdots = \frac{1}{1-\frac{1}{2}} = 2$$

in regions where both sides of eq. (E.3) are valid, but to apparently absurd results like

$$1 + 3 + 3^2 + 3^3 + \cdots = \frac{1}{1-3} = -\tfrac{1}{2}$$

in regions where the series expansion isn't valid. By comparison, eq. (E.1) seems tame.

Now back to business. The gamma function $\Gamma(s)$ is defined (for $\Re(s) > 0$) by

$$\Gamma(s) = \int_0^\infty dt\, e^{-t} t^{s-1} \tag{E.4}$$

An integration by parts shows that for integer values of s, $\Gamma(s)$ reproduces factorials:

$$\begin{aligned}
\Gamma(s+1) &= \int_0^\infty dt\, e^{-t} t^s \\
&= \int_0^\infty d(-e^{-t})\, t^s \\
&= -e^{-t}\, t^s \Big|_0^\infty - \int_0^\infty (-e^{-t})\, d(t^s) \\
&= 0 - \int_0^\infty (-e^{-t})\, s t^{s-1}\, dt \\
&= s \int_0^\infty dt\, e^{-t} t^{s-1} \\
&= s\Gamma(s)
\end{aligned}$$

In other words, for integer values of s, $\Gamma(s+1) = s!$.

With almost divine foresight, we will now consider

$$\int_0^\infty dt\, \frac{t^{s-1}}{e^t - 1}$$

If we multiply on top and bottom by e^{-t} and use the expansion (E.3), this becomes

$$\int_0^\infty dt\, \frac{t^{s-1}}{e^t - 1} = \int_0^\infty dt\, \frac{e^{-t} t^{s-1}}{1 - e^{-t}}$$

$$
\begin{aligned}
&= \int_0^\infty dt\, e^{-t} t^{s-1} \sum_{n=0}^{\infty} e^{-nt} \\
&= \sum_{n=0}^{\infty} \int_0^\infty dt\, e^{-(n+1)t} t^{s-1} \\
&= \sum_{n=1}^{\infty} \int_0^\infty dt\, e^{-nt} t^{s-1}
\end{aligned}
$$

which, with the change of variables $u = nt$, we can relate to the gamma function:

$$
\begin{aligned}
&= \sum_{n=1}^{\infty} \frac{1}{n^s} \int_0^\infty d(nt)\, e^{-nt} (nt)^{s-1} \\
&= \sum_{n=1}^{\infty} \frac{1}{n^s} \int_0^\infty du\, e^{-u} u^{s-1} \\
&= \sum_{n=1}^{\infty} \frac{1}{n^s} \Gamma(s)
\end{aligned}
$$

Since the factor $\Gamma(s)$ does not depend on n and can therefore be pulled outside the sum, we are left with the product of the zeta and gamma functions:

$$
\begin{aligned}
&= \Gamma(s) \sum_{n=1}^{\infty} \frac{1}{n^s} \\
&= \Gamma(s)\zeta(s)
\end{aligned}
$$

So far what we have is

$$
\int_0^\infty dt\, \frac{t^{s-1}}{e^t - 1} = \Gamma(s)\zeta(s) \tag{E.5}
$$

If we can analytically continue $\Gamma(s)$ and the integral on the left-hand side of eq. (E.5) into a region that includes $s = -1$, we can then extract from eq. (E.5) a result for $\zeta(-1)$. The trick is to recognize two things: first, that the badness in the integral on the left-hand side of eq. (E.5) and in the integral in the definition (E.4) of the gamma function are at the $t = 0$ end of the integrals, and, second, that these integrals from zero to infinity can be very neatly broken into integrals from zero to one and from one to infinity.

Let us first deal with the gamma function of eq. (E.4):

$$
\begin{aligned}
\Gamma(s) &= \int_0^\infty dt\, e^{-t} t^{s-1} \\
&= \int_0^1 dt\, e^{-t} t^{s-1} + \int_1^\infty dt\, e^{-t} t^{s-1}
\end{aligned}
$$

On the right-hand side the second integral is perfectly well behaved and in the first integral the badness is at the $t = 0$ limit: if we Taylor expand the e^{-t} in the first integral, we have

$$\Gamma(s) = \int_0^1 dt \left(1 - t + \tfrac{1}{2}t^2 + o(t^3)\right)t^{s-1} + \int_1^\infty dt\, e^{-t}t^{s-1} \tag{E.6}$$

where $o(t^3)$ stands for all the terms of order t^3, that is, terms that go as t^3 or higher powers of t. The first three terms in the expansion of e^{-t} thus contribute [2]

$$\begin{aligned}\int_0^1 dt\,(1 - t + \tfrac{1}{2}t^2)t^{s-1} &= \int_0^1 dt\,(t^{s-1} - t^s + \tfrac{1}{2}t^{s+1}) \\ &= \left[\frac{1}{s}t^s - \frac{1}{s+1}t^{s+1} + \frac{1}{2}\frac{1}{s+2}t^{s+2}\right]_0^1 \\ &= \frac{1}{s} - \frac{1}{s+1} + \frac{1}{2}\frac{1}{s+2}\end{aligned}$$

From this, we can see that the remaining terms in the expansion of e^{-t} will contribute terms that go as

$$\frac{1}{s+3}, \quad \frac{1}{s+4}, \quad \frac{1}{s+5}, \quad \cdots$$

The only term in $\Gamma(s)$ that is bad at $s = -1$ is the $-\frac{1}{s+1}$ term.

Next we deal with eq. (E.5) likewise:

$$\begin{aligned}\Gamma(s)\zeta(s) &= \int_0^\infty dt\,\frac{t^{s-1}}{e^t - 1} \\ &= \int_0^1 dt\,\frac{t^{s-1}}{e^t - 1} + \int_1^\infty dt\,\frac{t^{s-1}}{e^t - 1}\end{aligned} \tag{E.7}$$

Again, the second integral is perfectly well behaved and the badness is at the $t = 0$ limit of the first integral. This time, however, we need to Taylor

[2] The astute reader might object that the values of the coefficients of these three terms depend on our having chosen to break the integral from zero to infinity into integrals from zero to one and from one to infinity: had we chosen to break the integral into, say, integrals from zero to two and from two to infinity, then the contribution of the first these three terms would instead have been

$$\frac{1}{s}2^s - \frac{1}{s+1}2^{s+1} + \frac{1}{2}\frac{1}{s+2}2^{s+2}$$

For our present purposes, however, we will ultimately be comparing the expansions of the left- and right-hand sides of eq. (E.5), and for this comparison to be valid we must merely be consistent about how we break our integrals up.

expand $1/(e^t-1)$, which will be considerably more work than expanding e^{-t}. For small t, $e^t \approx 1+t$, so that

$$\frac{1}{e^t-1} \approx \frac{1}{t}$$

It will therefore be simpler to Taylor expand $t/(e^t-1)$, since $t/(e^t-1)$ is more polite at $t=0$. If we take

$$f(t) = \frac{t}{e^t-1}$$

then to obtain the explicit expansion

$$f(t) = f(0) + f'(0)\,t + \tfrac{1}{2}f''(0)\,t^2 + o(t^3) \tag{E.8}$$

we need to evaluate $f(0)$, $f'(0)$, and $f''(0)$. Since, as just noted, $1/(e^t-1) \sim 1/t$ for small t, we must evaluate these derivatives of $f(t)$ at $t=0$ with some care. For $f(0)$, we have

$$f(0) = \lim_{t\to 0}\frac{t}{e^t-1} = \lim_{t\to 0}\frac{t}{(1+t)-1} = \lim_{t\to 0}\frac{t}{t} = 1 \tag{E.9}$$

We can evaluate $f'(t)$ by the quotient rule:

$$\frac{d}{dt}\Big(\frac{t}{e^t-1}\Big) = \frac{(e^t-1)-te^t}{(e^t-1)^2}$$

Thus

$$f'(0) = \lim_{t\to 0}\frac{d}{dt}\Big(\frac{t}{e^t-1}\Big) = \lim_{t\to 0}\frac{(e^t-1)-te^t}{(e^t-1)^2}$$

Since, as we will find, the t^0 and t^1 terms in the numerator cancel out (as we would expect from the fact that the denominator goes as t^2), we need to keep terms out to t^2 in the numerator:

$$f'(0) = \lim_{t\to 0}\frac{\Big((1+t+\frac{1}{2}t^2)-1\Big)-t(1+t)}{t^2} = \lim_{t\to 0}\frac{-\frac{1}{2}t^2}{t^2} = -\tfrac{1}{2} \tag{E.10}$$

Finally, we evaluate $f''(t)$ by using the quotient rule on our result for $f'(t)$:

$$\begin{aligned}\frac{d^2}{dt^2}\Big(\frac{t}{e^t-1}\Big) &= \frac{d}{dt}\left(\frac{e^t-1-te^t}{(e^t-1)^2}\right)\\ &= \frac{(e^t-1)^2(-te^t)-(e^t-1-te^t)\cdot 2(e^t-1)}{(e^t-1)^4}\end{aligned}$$

which, if we divide out one factor of $e^t - 1$, simplifies to

$$= \frac{-(e^t-1)te^t - 2(e^t - 1 - te^t)}{(e^t-1)^3}$$

This time, we need to keep terms out to t^3 in the numerator when we evaluate $f''(0)$:

$$\begin{aligned}
f''(0) &= \lim_{t\to 0} \frac{-(e^t-1)te^t - 2(e^t - 1 - te^t)}{(e^t-1)^3} \\
&\qquad -\Big((1+t+\tfrac{1}{2}t^2)-1\Big)t(1+t) \\
&= \lim_{t\to 0} \frac{\qquad -2\Big((1+t+\tfrac{1}{2}t^2+\tfrac{1}{6}t^3)-1-t(1+t+\tfrac{1}{2}t^2)\Big)(1+t)}{t^3} \\
&= \lim_{t\to 0} \frac{-(t+\frac{1}{2}t^2)t(1+t) - 2(-\frac{1}{2}t^2 - \frac{1}{3}t^3)(1+t)}{t^3} \\
&= \lim_{t\to 0} \frac{-(t^2+\frac{3}{2}t^3) + 2(\frac{1}{2}t^2 + \frac{5}{6}t^3)}{t^3} \\
&= \lim_{t\to 0} \frac{\frac{1}{6}t^3}{t^3} \\
&= \tfrac{1}{6}
\end{aligned} \tag{E.11}$$

Using our results (E.9) through (E.11) in eq. (E.8), we have

$$\frac{t}{e^t-1} = 1 - \tfrac{1}{2}t + \tfrac{1}{2}\tfrac{1}{6}t^2 + o(t^3) = 1 - \tfrac{1}{2}t + \tfrac{1}{12}t^2 + o(t^3)$$

And so

$$\frac{1}{e^t-1} = \frac{1}{t} - \frac{1}{2} + \frac{1}{12}t + o(t^2)$$

Using this expansion in eq. (E.7), we can isolate the badness in $\Gamma(s)\zeta(s)$ at $s = -1$:

$$\begin{aligned}
\Gamma(s)\zeta(s) &= \int_0^1 dt\, \frac{t^{s-1}}{e^t-1} + \int_1^\infty dt\, \frac{t^{s-1}}{e^t-1} \\
&= \int_0^1 dt\, t^{s-1}\Big(\frac{1}{t} - \frac{1}{2} + \frac{1}{12}t + o(t^2)\Big) + \int_1^\infty dt\, \frac{t^{s-1}}{e^t-1} \\
&= \int_0^1 dt\, \Big(t^{s-2} - \tfrac{1}{2}t^{s-1} + \tfrac{1}{12}t^s + o(t^{s+1})\Big) + \int_1^\infty dt\, \frac{t^{s-1}}{e^t-1} \\
&= \Big[\frac{1}{s-1}t^{s-1} - \frac{1}{2}\frac{1}{s}t^s + \frac{1}{12}\frac{1}{s+1}t^{s+1}\Big]_0^1
\end{aligned}$$

$$+\int_0^1 dt\, o(t^{s+1}) + \int_1^\infty dt\, \frac{t^{s-1}}{e^t-1}$$
$$= \frac{1}{s-1} - \frac{1}{2}\frac{1}{s} + \frac{1}{12}\frac{1}{s+1} + \int_0^1 dt\, o(t^{s+1}) + \int_1^\infty dt\, \frac{t^{s-1}}{e^t-1}$$

with the contributions from the integral of $o(t^{s+1})$ giving terms that go as

$$\frac{1}{s+2}, \quad \frac{1}{s+3}, \quad \frac{1}{s+4}, \quad \cdots$$

The badness in $\Gamma(s)\zeta(s)$ at $s=-1$ is therefore all in the $\frac{1}{12}\frac{1}{s+1}$ term.

At this point, we have ascertained that as $s \to -1$ the terms to $\Gamma(s)$ and $\Gamma(s)\zeta(s)$ that blow up are $-\frac{1}{s+1}$ and $\frac{1}{12}\frac{1}{s+1}$, respectively. Near $s=-1$, these terms will become arbitrarily large and we can therefore separately equate their contributions in the limit $s \to -1$. Moreover, since $\Gamma(s)$ and $\Gamma(s)\zeta(s)$ both blow up as $1/(s+1)$ near $s=-1$ (as opposed to one blowing up as $1/(s+1)^p$ and the other as $1/(s+1)^q$ with $p \neq q$), $\zeta(s)$ must be well behaved as $s \to -1$. Thus

$$\lim_{s\to-1} \Gamma(s)\zeta(s) = \lim_{s\to-1} \int_0^\infty dt\, \frac{t^{s-1}}{e^t-1}$$
$$\lim_{s\to-1} -\frac{1}{s+1}\zeta(-1) = \lim_{s\to-1} \frac{1}{12}\frac{1}{s+1}$$
$$-\zeta(-1) = \tfrac{1}{12}$$

so that $\zeta(-1) = -\frac{1}{12}$ and we have

$$\zeta(-1) = \sum_{n=1}^\infty \frac{1}{n^{-1}} = \sum_{n=1}^\infty n = 1+2+3+4+\cdots = -\tfrac{1}{12}$$

Appendix F

And Now Some Completely Gratuitous Pictures of Field Lines

. . . to annoy the censors and hopefully spark some sort of controversy, which seems the only way these days of getting the jaded, video-setted public off their ******* ***** and back in the ******* cinema.[1]

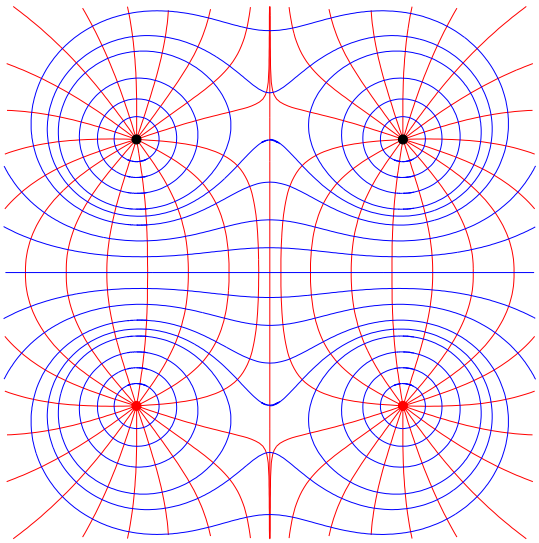


Figure F.1: Field & Equipotential Lines of Point Charges

Fig. (F.1) shows the field and equipotential lines of four charges arranged at the corners of a square, with two $+1$ charges on top and two -1 charges on the bottom.

[1] From the end of Monty Python's *The Meaning of Life*. Though we may not have the quote down exactly.

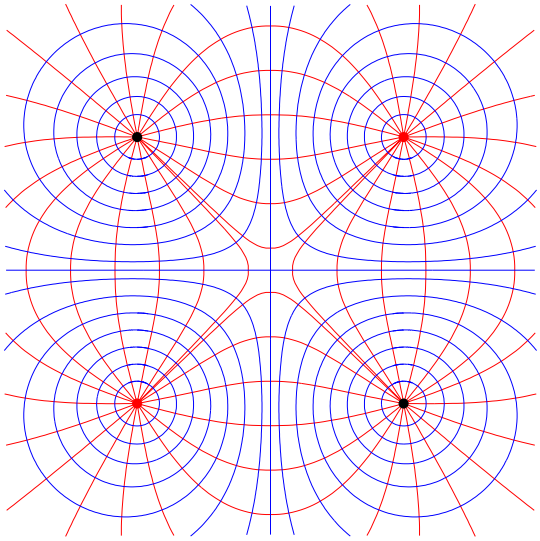


Figure F.2: More Field & Equipotential Lines of Point Charges

Fig. (F.2) shows the field and equipotential lines of a similar arrangement of charge, but with the two positive charges along one diagonal and the two negative charges along the other.

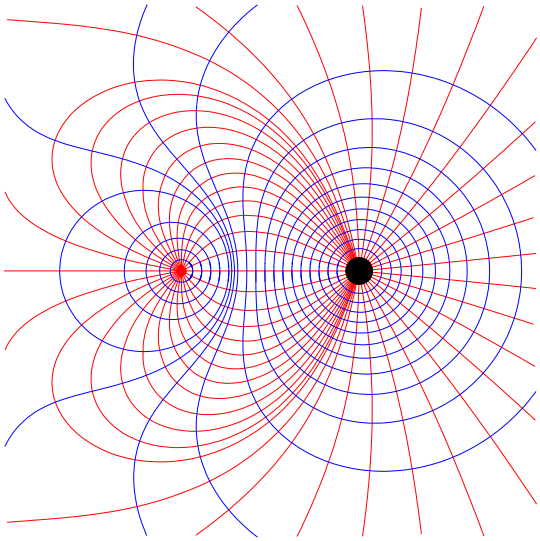


Figure F.3: Still More Field & Equipotential Lines of Point Charges

Fig. (F.3) shows the field and equipotential lines of a $+1$ charge on the left and a -3 charge on the right.

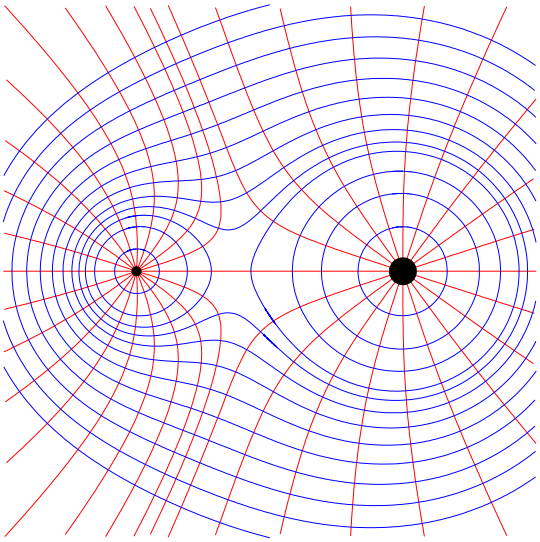


Figure F.4: Field & Equipotential Lines of Point Charges Yet Again

Fig. (F.4) shows the field and equipotential lines of a similar arrangement of charge, but with both charges of the same sign.

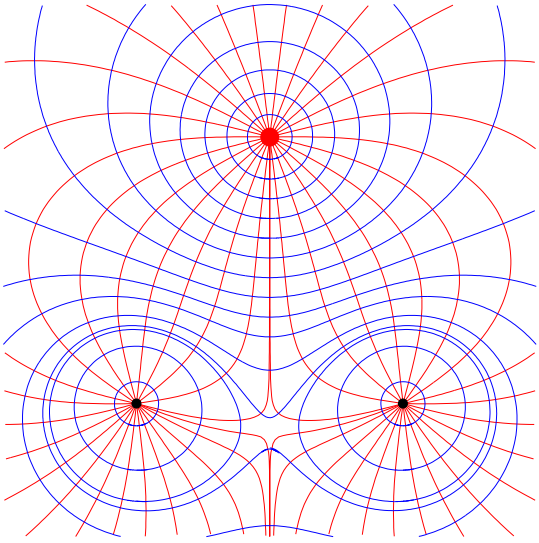


Figure F.5: Field & Equipotential Lines of Point Charges One Last Time

Fig. (F.5) shows the field and equipotential lines of a $+2$ charge above two -1 charges.

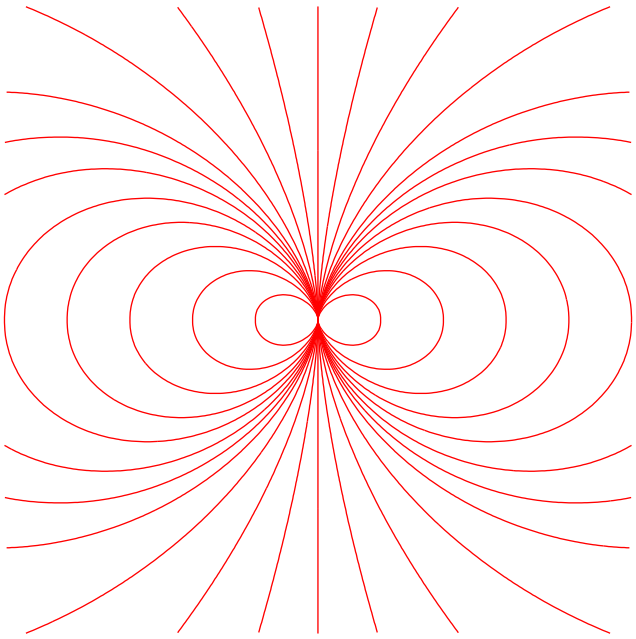


Figure F.6: Field Lines Far from an Electric Dipole

Fig. (F.6) shows the field lines of an electric dipole that consists of a pair of equal and opposite charges (say, $+1$ and -1). In fig. (F.6) we are far enough away that the two charges seem to be at a single point.

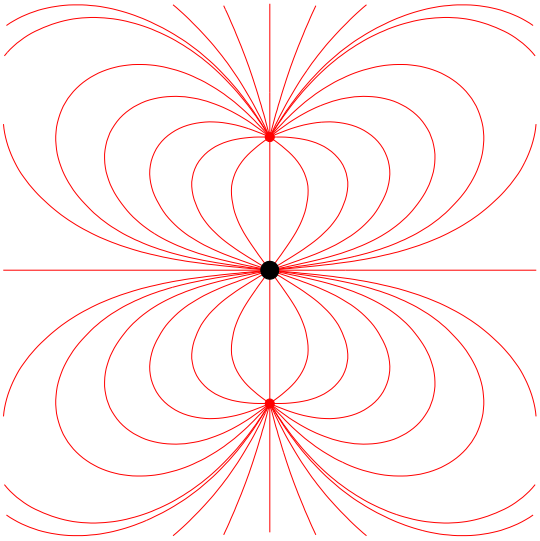


Figure F.7: Field Lines of an Electric Quadrupole

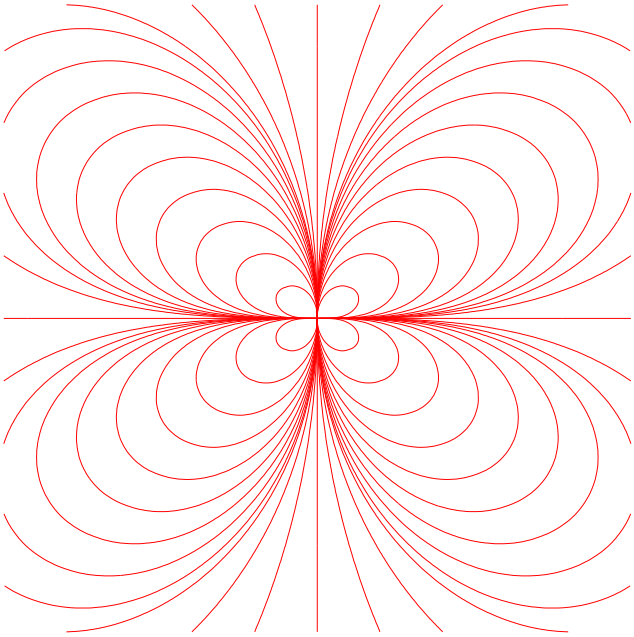


Figure F.8: Field Lines Far from an Electric Quadrupole

Figg. (F.7) and (F.8) show the field lines of an electric quadrupole, in which the charges are proportionally $+1$, -2, and $+1$. In Fig. (F.7) we are close enough to the charges to see the separation between them; in fig. (F.8) we are far enough away that the charges seem to be at a single point.

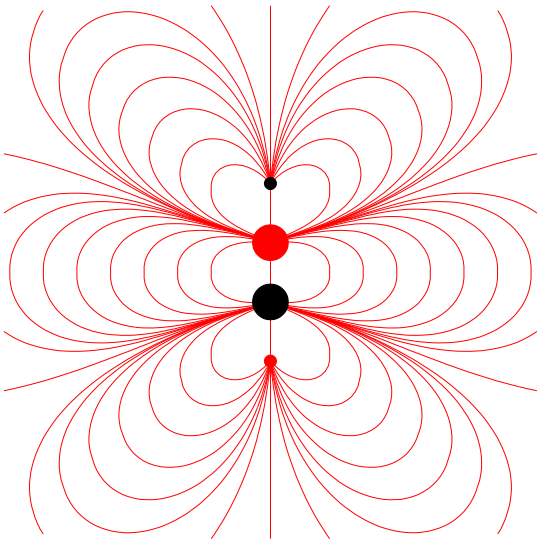


Figure F.9: Field Lines of an Electric Octopole

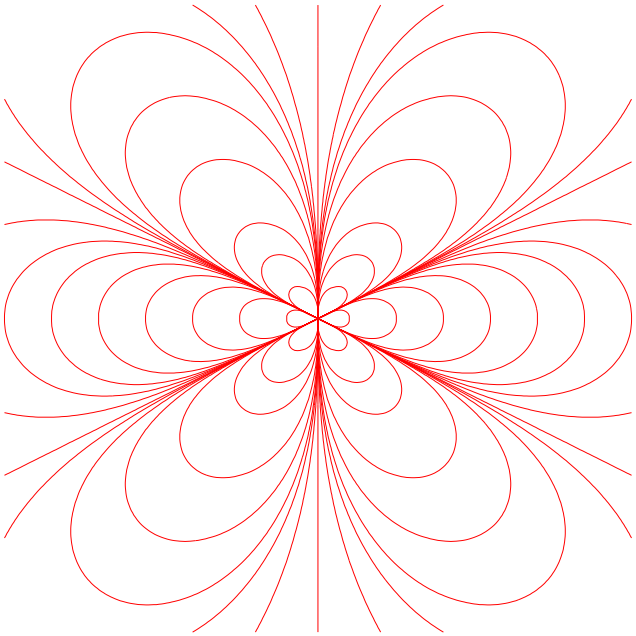


Figure F.10: Field Lines Far from an Electric Octopole

Figg. (F.9) and (F.10) show the field lines of an electric octopole, in which the charges are proportionally -1, $+3$, -3, and $+1$. In Fig. (F.9) we are close enough to the charges to see the separation between them; in fig. (F.10) we are far enough away that the charges seem to be at a single point.

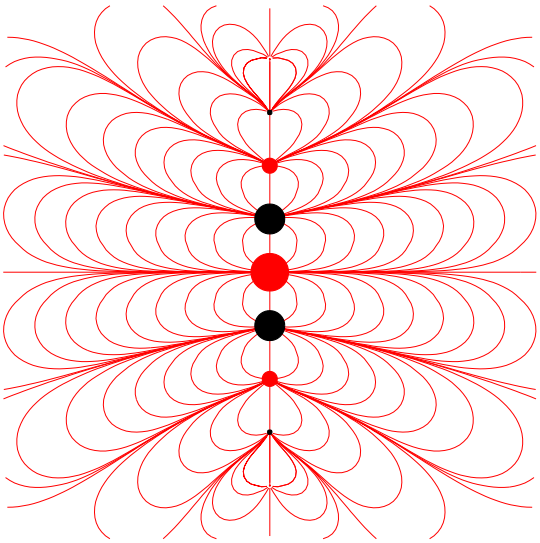


Figure F.11: Field Lines of an Electric 256-pole

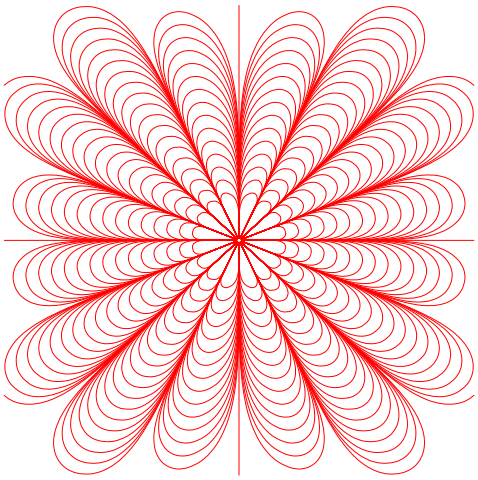


Figure F.12: Field Lines Far from an Electric 256-pole

Figg. (F.11) and (F.12) show the field lines of an electric 256-pole, which consists of a chain of nine charges in the proportions $+1$, -8, $+28$, -56, $+70$, -56, $+28$, -8, and $+1$. In Fig. (F.11) we are close enough to the charges to see the separation between them; in fig. (F.12) we are far enough away that the charges seem to be at a single point.

Afterword

If you are reading this, you exist in the present. And for a short while you will continue to exist in the future. But, as so poetically pointed out by Kurt Vonnegut in *Slaughterhouse-Five*, you will *always* exist in the past.

Of course, fat lot of consolation that will be.

Mais à quelle fin ce monde a-t-il donc été formé? dit Candide.

Pour nous faire enrager, réspondit Martin

...

Il faut cultiver notre jardin.

— Voltaire, *Candide*.

Vanity of vanities, saith the Preacher, vanity of vanities; all is vanity. ... The thing that hath been, it is that which shall be; and that which is done is that which shall be done: and there is no new thing under the sun. ... There is no remembrance of former things; neither shall there be any remembrance of things that are to come with those that shall come after. ... Wherefore I perceive that there is nothing better, than that a man should rejoice in his own works; for that is his portion: for who shall bring him to see what shall be after him? ... Behold that which I have seen: it is good and comely for one to eat and to drink, and to enjoy the good of all his labor that he taketh under the sun all the days of his life, which God giveth him, and to have a few good cigars before they turn out the lights: for it is his portion.

— *Ecclesiastes* 1.2, 9, 11; 3.22; & 5.18. Though the part about the cigars is ours. And a fine interpolation it is, too, if we say so ourselves. The original author would doubtless have included it himself if he'd thought of it.

A Bibliography of Sorts

You are born. Then for a long period of time nothing makes any sense at all. Then you die.

We are all going to die. And to add insult to injury, we will die without ever really understanding why the chicken crossed the road.

But while you're here, at least you can read a few good books.

For what it's worth, we recommend the following if you want to learn some more physics: [2]

Books About Physics

Brian Greene, *The Elegant Universe*. Greene gives a generally excellent discussion of all of physics, including a sort of pseudohistory, and emphasizing string theory and the corresponding cosmology. He arguably goes a bit over the top with the epic melodrama of his description of the discord between quantum theory and general relativity, and he gives short shrift to the hugely important topic of quantum field theory, but the book is otherwise a magnificent work. If you are only going to read one book about modern physics, this should be it.

Brian Greene, *The Fabric of the Cosmos*. In this very thorough and excellent exposition, Greene traces the historical development of our understanding of space and time from Newton through string and M theory, including discussions of the arrow of time and of cosmology. While, because of its emphasis on the theme of space and time, the more angular perspective of *The Fabric of the Cosmos* does not provide as general an introduction to our current understanding of the physical universe as *The Elegant Universe*, this book is arguably the second book you should read if you are only going to read two books.

[2] Usually bibliographic entries cite the publisher, year of publication, etc., but we aren't here to do homage to a system of corporate greed.

Brian Greene, *The Hidden Reality*. In yet another thorough and excellent exposition, Greene discusses the amazing consequences of the various cosmological possibilities, from the classic infinite universe to the multiverse scenarios of eternal inflation, string theory, and other modern theories. While necessarily abbreviated in its presentation of the background physics, the text is nonetheless very lucid and has a tremendous "wow factor." If you are only going to read three books about modern physics, this should be the third.

Leon Lederman, *The God Particle*. An ardent, even chauvinistic, experimentalist, Lederman openly biases his presentation toward experimental history and results, but he nonetheless does an excellent job of explaining collider/accelerator physics. His outline of the physics of the ancient Greeks is very good, and the whole book is written with an entertaining sense of humor. The title refers to the search for the Higgs particle.

Alan Guth, *The Inflationary Universe*. Guth gives a thorough and very lucid presentation of modern cosmology, with emphasis on inflation, a theory that he developed and arguably remains the foremost expert on. The book also has some interesting personal history that gives you an idea what it is like to be a theoretical physicist.

Lisa Randall, *Warped Passages: Unraveling the Mysteries of the Universe's Hidden Dimensions*. Because it restricts itself largely to those background aspects of relativity and quantum theory directly relevant to its titular thesis, *Warped Passages* lacks some of the "wow factor" of *The Elegant Universe*, but it makes an excellent followup to Greene for those interested in a more thorough discussion of technical aspects of the physics and of extra dimensions in particular. You can skip over the silly story with which each chapter begins; the rest is a remarkably clear and worthwhile exposition.

Weinberg, *The First Three Minutes*. Weinberg's book is still quite valid in spite of its vintage (1977), and recent editions have an afterword with some discussion of what has been learned since it was first published. While his presentation might be a bit more technical than the average pedestrian would find palatable, Weinberg gives an excellent chronology of standard big-bang cosmology.

Real Physics Books

Arfken & Weber, *Mathematical Methods for Physicists*. We begin by listing a math text because at this stage the principal impediment to your tackling more advanced physics is your limited knowledge of math. Arfken is

arguably more a reference than a textbook, and (at least in the fifth edition) the mathematical typesetting is horrible and the cross-referencing and index sometimes seem to be for a different edition. But the book is neatly organized and generally quite lucid; it has worked examples and exercises for each section; and if the section in Arfken on the particular aspect of math of interest to you doesn't meet your needs, usually it will at least clue you in to exactly what you need to learn, so that you can search for it more effectively in other texts.

Feynman et al., *The Feynman Lectures on Physics*. An excellent introductory opus.

The Berkeley Physics Course. An excellent five-volume series on introductory physics.

Symon, *Mechanics*. Not ideal, but a decent undergraduate text on classical mechanics. If we had to pick a single undergraduate text on classical mechanics, this would be it. And since we have in fact just picked a single undergraduate text on classical mechanics, this was it.

Goldstein, *Classical Mechanics*. The definitive graduate-level work on classical mechanics.

Jackson, *Classical Electrodynamics*. The definitive graduate-level work on classical electrodynamics.[3]

Taylor & Wheeler, *Spacetime Physics*. An introductory and yet very thorough treatment of relativity, with many worked problems. The book uses a minimum of math and will be more than accessible at your current level.

Albert Einstein, *The Meaning of Relativity*. A very physical treatment of relativity, as The Man himself saw it. If you can get a handle on four-vector notation and the summation convention, the parts on special relativity will be accessible at your current level, and the earlier parts on general relativity might even be within reach.

Steven Weinberg, *Gravitation & Cosmology*. One of the two most physical and accessible of the classic pedagogical texts on general relativity. Weinberg, as is his wont, is all business; the presentation is very direct and to the point. At the late undergraduate and graduate level.

Misner, Thorne, & Wheeler, *Gravitation*. The other of the two most physical and accessible of the classic pedagogical texts on general relativity. At 1200-odd pages, the book is a very thorough introduction to general relativity.

[3]We wish we could also list an undergraduate-level text, but we are unaware of a really decent one among the myriads out there.

Although most of the text would have to be considered graduate-level, much of the first chapter is accessible at your current level and may be interesting to you.

Gasiorowicz, *Quantum Physics*. One of the very few good introductory books on quantum mechanics. At the level of undergraduate seniors.

Irene V. Schensted, *On the Applications of Group Theory to Quantum Mechanics*. An excellent introduction to group theory generally, and in particular to the Lie-group symmetries that are so fundamentally important in physics. The earlier parts of the book should be accessible at your current level. Unfortunately the book is very difficult to find — written and privately printed (with a painfully funky typesetting) by what seems to have been a hippie mathematician up in Maine, it was never widely circulated.[4]

Itzykson & Zuber, *Quantum Field Theory*. Although now somewhat dated, probably still the best pedagogical text on quantum field theory. It suffers from typos, and a reader seeking a first introduction to quantum field theory will likely find it less than ideal, but it doesn't have much competition for thoroughness, lucidity, or number of worked examples.

Steven Weinberg, *Quantum Field Theory*, in three volumes. A wonderfully thorough, detailed, and insightful text on field theory, but unfortunately accessible only if you already know the subject.

Michio Kaku, *Quantum Field Theory*. Probably not the best text if you are learning field theory for the first time, but notable for taking the more modern approach of developing quantum field theory from symmetry principles.

Feynman & Hibbs, *Quantum Mechanics & Path Integrals*. An introduction to the path integral formalism from the person (Feynman) who developed it. For those of you with a particular interest in path integrals.

Barton Zwiebach, *A First Course in String Theory*. An excellent undergraduate introduction to string theory, remarkable for its lucidity. The earlier chapters will already be accessible to you.

Polchinski, *String Theory*, in two volumes. A very thorough pedagogical introduction to string theory at the graduate level.

Donald Marolf, `https://arxiv.org/abs/hep-th/0311044`, is a nice guide to the literature of string theory, from the popular level to the real thing.

[4]Yet another sad example of the way our culture of corporate greed tramples the virtuous and beautiful in its unprincipled pursuit of profit. Contrary to the prevailing myth, what private enterprise and free-market "competition" actually produce is a materialistic mediocrity that nurtures the basest sort of lowest-common-denominator culture.

Benediction

To paraphrase Dustin Hoffman's character at the end of the grade-B movie *The Hero*:

> Everything is b*******. But there's, like, different levels of b*******. So you gotta find the level of b******* you feel comfortable with, and that's your b*******!

And so we leave you with this blessing:

May you find the level of b******* with which you feel comfortable.

Index

Symbols

3° background 1105
α decay 639
α particle 639
β (relativity)
  definition 534
β decay 639–641
β particle 639
β^+ decay 640
β^- decay 640
$\gamma = C_p/C_V$ 699, 703
γ (relativity)
  definition 534
γ decay 641
γ particle 641
δ function
    see delta function
Ω (unit) 841

A

absolute temperature 671
absolute zero 671
absorbed dose (radiation) 649
AC 868, 969–987
  treating as DC 868–870
acceleration 123, 126
  angular
    see angular acceleration
  average 127
  centripetal 142, 144, 148
  circular tangential 147
  constant 132–135, 588–598
  Coriolis 152
  gravitational 195, 209, 476
  instantaneous 127
  interpretation of direction 128–129
action
  definition 1005
  principle of least
    see principle of least action
action at a distance 754, 759, 765, 1039–1041, 1043
activity 643
addition of velocities
    see relativity, addition of velocities
adiabatic process
    see thermal process, adiabatic
Æsop 605
æther 184, 492, 514, 517, 1040
affine connection 585
air resistance 1113–1132
Ampère (unit) 751
Ampère's law 758–760, 890, 892–901, 927–928
  equivalence of differential and integral forms 759–760
  normal, convention for direction of 760
Amperian loop 893, 894, 897
amplitude 459
  wave *see* wave amplitude
angular
  acceleration 139
  frequency 143, 459
  momentum 353–354, 360–362
    center of mass, about 383
    center of mass, of 383
    conservation 381–390
    definition 353, 360
  velocity 139, 143
anharmonic oscillations 485
antineutrino 1069
Archimedes's principle 623–625
Aspect, Alain 1044
astigmatism 45
atmosphere (unit) 622, 678
atomic mass 638
atomic mass unit 608, 638
atomic number 638

B

ballistic pendulum
    see pendulum, ballistic
banking 207–208
baryon 1070
batteries
  in parallel 880
  in series 880
Becquerel (unit) 643
Bell's inequalities 1044
Bell, John 1044
Bernoulli equation 621–623
binding energy 638
biological effects of radiation
    see radiation, biological effects
Biot, Jean-Baptiste 901
Biot-Savart law 901–910
black hole 583, 586, 587
blackbody radiation 676, 1050
body, extended
    see extended body
bogosity 197, 254, 263, 444, 784, 841
Bokonon 49, 226, 330
Boltzmann constant 666, 670, 677, 726, 730
  relation to gas constant 677
Boltzmann factor 666–668, 731
Boltzmann, Ludwig 666
boson 1096–1097
boundary conditions 465, 467–468
brachistochrone
  derivation 997–1001
  properties 1002–1004
braking, magnetic 947
British thermal unit 716
BTU (unit) 716

C

calculus of variations 993–996
calorie (unit) 673–674
capacitance 851–868
  coefficients of 852
  equivalent
    parallel 859–860, 877
    series 859–860, 877
capacitor 799, 852
  cylindrical 853–854
  electrolytic 856
  energy stored in 861–862
  flux 852
  games with 862–865
  impedance 979–980
  parallel-plate 852–853
  spherical 854–855
  variable 987
carbon dating 659
Carnot cycle 704–717
  description of 705–706
  efficiency 713, 716–717
  fr33 pics 4 u 708
  plot of 705
  reversibility 718–719
Carnot, N.L.S. 704
catenary 216–220
Celsius scale 670
center of gravity
    see gravity, center of
center of mass 193, 291–297
  acceleration 298
  definition 291, 292
  holes, dealing with 297
  superposition 296–297
  velocity 298
center-of-mass and relative coordinates 308
center-of-mass frame 303–306, 308–311
centigrade scale 670
central force 1008, 1141
centrifugal potential
    see potential, centrifugal
centripetal whatnot
    see whatnot, centripetal
chain reaction 644
charge 750–751
  conjugation 1072
  density 751, 776–777, 839
  fundamental 637, 815
circuit 839, 842–868
  analysis 846–851
  diagrams 842–845
  parallel connection
    capacitors 859
    resistors 844–846
  power in 842
  series connection
    capacitors 859–860
    resistors 844–845
  symbol

capacitor 858
inductor 955
resistor 842
voltage 842
symmetry in 848
circular motion
see motion, circular
cloud chamber 916
co-dependence, dysfunctional 426, 429
coaxial cable 921
cold fusion 646–647, 1059
Coleman-Mandula theorem 1096
college admissions 7, 1038
collision 302–303
completely inelastic 303, 306, 311
elastic 303
one-dimensional two-body 311–314
inelastic 303
two-body 307–314
color $SU(3)$ 1070
color vision 31–32
common spacetime origin 539
commutator 1061
complex-arithmetic refresher 974–976
conductivity 812, 840
conductor 809–812, 840
charge in/on 810, 811
circuits, use in 811
electric field at surface 811
electric field within 809
equipotentiality 810
confinement 1073
conjugate momentum 1007
constraints 1011–1019
forces of 1015–1019
continuity, equation of 623, 962
conversion factors 19–21
coordinates
center-of-mass and relative 308
cyclic 1008
generalized 1005
ignorable 1008
Coriolis
effects 366–373
force 371
cosmological constant 579, 585, 1107, 1109, 1110
cosmology 1102–1110
couch potato 260
Coulomb (unit) 751
Coulomb's semibogus law 784–786
coupling constant 1075
critical angle 34
critical density 1107–1109
critical mass 644
cross product 71, 73–78
crystal meth 705
Curie (unit) 643, 659
curl 100, 102, 104–106
current 751–753, 839–851
density 751, 752, 839
electric field, relation to 840
eddy 947
induced 932–934, 936
magnetostatic restrictions on 890
curve balls 626–627
cycle 143
thermodynamic
see thermal cycle
cyclic coordinates
see coordinates, cyclic
cyclops 176, 223
plural 176
cyclotron frequency 915

D

damping, magnetic 947
DC 868
de Broglie, Louis 1051
dead horses, beating 269–271
death 3, 393, 1173
decay constant 642
relation to half-life 643
decibels 489
table of values 490
decoherence 1045
degree of freedom 684, 698, 703, 1007, 1011–1012
delta function 970–972, 987–990
density, charge
see charge density
der-der tube 897
deuterium 645
dielectric 856–858
constant 858
diffraction 48
diode 944

diopter 45
dipole 756
  electric 797–799, 822–823, 825, 912, 1166
  magnetic 756, 910–914
  moment
    electric 797
    magnetic 912
Dirac equation 1078–1093
dispersion 48–49
displacement 124
  vs distance 124
divergence 96, 98, 104–106
Doppler shift
  light 491, 577–579
  sound 490–492
dosimetry 648–651
dot product 71–73, 78
double-speak 674

E

Earnshaw's theorem 826
Easter Bunny 722, 852
economics 263, 1038
Einstein 511, 517, 518, 563, 570, 579–580, 585, 749, 768, 1041, 1043–1044, 1050, 1094
  awesomely righteous dude 516
  weaselly wascal 1043
electric *blah blah*
  see blah blah, electric
electrodynamics 927–943
electromagnetic interaction 1069, 1071, 1073, 1075, 1077, 1093, 1094, 1103, 1105
electromagnetism
  parity (chirality) of 920
  relativity and 768–772, 942
electromotive force
  see EMF
electron capture 640
electron volt 670
electron volt (unit) 608
electroweak interaction 1071, 1072, 1077, 1095, 1103
EMF 932–935, 937
  versus voltage 932–933
energy 253–271
  conservation 253–254, 261, 1135–1140
  internal
    see internal energy
  kinetic 255
    center-of-mass 305, 306, 310
    center-of-mass vs relative 310–311
    definition 255
    relative 305, 310
    relativistic 573
    relativistic vs Newtonian 535
    rotational 349–350, 358–360
    total 305, 310
    translational vs rotational 390–392
  operator 1052
  potential 260–264
    definition 261
    electrostatic 807–809
    electrostatic, relation to electrostatic potential 807
    gravitational 266–268
    normal force 266
    spring 269
  relativistic 571–576
  total, definition 261
energy-momentum tensor 585, 1107, 1135
entropy
  coiffure and 732
  Darwin and 732–733
  ideal gas 729–731
  statistical mechanical 725–732
    relation for 726
  thermodynamic 719–725
    relation for 720
episode, psychotic 965
EPR paradox 1043
equation of continuity
  see continuity, equation of
equation of state
  see state, equation of
equilibrium
  point 287
  saddle point 447
  stable 287, 445–447, 464
  static 441–447
    conditions 441
    torque theorem 444–445
  unstable 287, 445–447
equilibrium constant 669
equipotential 802–807

equipotential lines 802–807, 1163–1169
equivalence, principle of
see principle of equivalence
equivalent dose (radiation) 649
escape velocity
see velocity, escape
estimates 21–24
Euler-Lagrange equation
see Lagrange equation
eV (unit) 608
event horizon 564
Everett, Hugh 1044
evil, love of 4
existence, futility of
see life, futility of
extended body 193, 347
extensive variable 666
eye
shooting out 312
vision 43–45

F

Fahrenheit scale 670
Fahrenheit, Daniel Gabriel 670
false vacuum 1104
Farad 852
Faraday's law 756–758, 927–928
equivalence of differential and integral forms 756–758
normal, convention for direction of 758
farsightedness 44
fermion 1095–1097
ferromagnetism 913
Feynman diagrams 1074–1075, 1092–1093
Feynman, Richard 1046
field
electric 753–755, 773–799
calculating from electrostatic potential 801
determining by Gauss's law 773–783
direct integration of 786–794
discontinuity across surface charge 790–791
energy density 862, 958–962
induced 928, 931
relation to electrostatic potential 799
magnetic 755–756
determining by Ampère's law 893–901
determining by Biot-Savart law 905–910
energy density 955, 958–962
right-hand rule for 896, 900
solenoidal 897–901, 909–910
field lines, electric 794–797, 1163–1169
first law of thermodynamics
see thermodynamics, first law
fission, nuclear 644–645, 647
flavor 1070
flow rate 623
fluid dynamics 621–628
flux
electric 759–760
magnetic 757–758, 927, 928, 939–942
food calorie (unit) 673–674
force
central 117, 1008
centrifugal vs centripetal 205–206, 500
centripetal 204–205
conservative 262
diagrams 197–203
electric 753–754
fictitious 206, 368, 371, 500, 513
gravitational *see* gravity
magnetic 759, 891–892
on a wire 892
properties of 891
normal *see* normal force
relativistic 575
spring *see* spring force
force of constraint
see constraints, forces of
four-acceleration 591–593
four-vector 570, 583, 589–593, 1079, 1135, 1149
dot product 591
four-velocity 589–591
Fourier transform 467, 969–974
definition 970, 971
frame
see reference frame
free-body diagram

see force diagram
free-fall 135
freedom, degree of
see degree of freedom
frequency 143, 459
angular
see angular frequency
friction 195, 196
almost nonbogus 215–216, 1113
bogus 196
coefficients of 197
kinetic 196
potential energy and 262–263
semibogus 214–215, 1113–1132
static 196, 197
Friedmann equations 1108
Friedmann, Alexander 578
fringe field 921, 951
fundamental (standing wave) 487
fundamental charge
see charge, fundamental
fusion, nuclear 644–648

G

G 208, 1107
g vs G 209
Galileo 1038–1039
galvanometer 880
gamma matrix 1081
gas constant 677
relation to Boltzmann constant 677
gauge
field 1085
symmetry
see symmetry, gauge
transform 765, 902, 1085
Gauss's law 753–755, 773–783
equivalence of differential and integral forms
754–755
magnetic 755–756
equivalence of differential and integral forms
755
Gauss's theorem 96–100
Gaussian surface 755, 773, 774, 777, 778, 781, 811
gedanken experiment 558
general coordinate transform 586
general relativity
see relativity, general
generalized blah blah
see blah blah, generalized
generator, electric 937–939
homopolar 948
geodesic 581
gerrymander 254, 256, 298, 392, 499, 529, 538, 568, 574, 729, 754, 757, 770, 892, 959, 995, 1062, 1063, 1137, 1145, 1147
gluon 1070–1071, 1093
governor 239, 431
gradient 94–95
grand unified theories
1095–1096
gravitational bending of light
581
gravitational field equations
see relativity, general, field equations
gravitational interaction 1071, 1073, 1076, 1094, 1103
gravitational slingshot 314
graviton 1071
gravity
center of 324, 393
near Earth's surface 209, 266–267
Newtonian 208–212, 267–268, 393, 580, 754, 1039, 1041
miscellaneous properties 317–323
spheres and shells 317–321, 784–785
quantum 586, 1072, 1076, 1077, 1096
relativistic
see relativity, general
Gray (unit) 650
gremlins, invisible 48
Guth, Alan 1105
gyroscopic motion 401–403, 628

H

hadron 1070
half-life 641, 642
relation to decay constant 643
Hamiltonian 1059
hare, the 605
harmonic oscillations
see motion, harmonic
harmonic oscillator

classical
see motion, harmonic
quantum 269, 1059–1064, 1066
thermal 684–685
heat capacity 698, 703
heat engine 693, 704
efficiency 713
heat pump 740
Heaviside, Oliver 1039
heavy water 645
Heisenberg uncertainty principles
see uncertainty principles
Helmholtz's theorem 112–114
Henry (unit) 955
Hermitian conjugate 1060, 1081
Hertz (unit) 143
hidden-variable theories 1044
Higgs mechanism 1072
Higgs particle 579, 1029, 1072, 1077
high-pass filter 983
Hilbert, David 518
Hokey-Pokey, the 1035
homopolar generator 948
horizon
see event horizon
Hubble constant 1108
Hubble, Edwin 579
hydrogen atom 1053–1054
hyperopia 44

I

ideal gas 676
ideal-gas law 676–684
ignorable coordinates
see coordinates, ignorable
ignorance, blissful 253
image
height relative to object 39, 40
real 36
virtual 36
images, method of
see method of images
imagination, crushing into dust 463
impedance 978–980
values of 980
impulse 334
inclines 196
inductance 953–958
EMF across 954
equivalent 963
flux, relation to 953
mutual 953
induction
electric 759
magnetic 756
inductor 953
energy stored in 954–955
impedance 979–980
inertia tensor 358, 359
inertia, moment of
see moment of inertia
inflation 579, 1035, 1103–1105, 1109–1110
eternal 1100, 1105
inflaton 579, 1104, 1110
integers, sum of positive 555, 1075, 1098, 1155–1161
integration
line 87–90
surface 92
volume 92–93
intensive variable 666
internal energy 684–687, 698, 702
interval, invariant
see invariant interval
invariant interval 530, 563
irreversibility
see reversibility
isobaric process
see thermal process, isobaric
isochoric process
see thermal process, isochoric
isochrone 1002–1004
isothermal process
see thermal process, isothermal
isotope 638

J

JB, the legendary 601–602
Joule (unit) 256

K

Kaluza, Theodor 1094
Kaluza-Klein theory 1094–1095
karma, good 5–7
Kelvin scale 670–671
Kepler's laws 209–212, 1039, 1041, 1141–1148

kinematics, definition 123
kinetic theory 678
Kirchhoff's rules
see loop and junction rules
Klein, Oskar 1094

L

lab frame 307
Lagrange
equation 1007
multipliers 1011–1019
Lagrangian
definition 1005
density 1029, 1081, 1136
dynamics 1004–1019
Laplacian 110
latent heat 673
length contraction
see relativity, length contraction
lens 37–42
concave vs convex 42
converging 41
corrective 43–45
diverging 41–42
equation 38–40
eyepiece 42
focal length 37
foci 37
objective 42
rules 38
Lenz's law 932, 933, 935
lepton 1069, 1093
lies and deceit 191, 197, 208, 211, 213, 843, 1099
life, futility of
see existence, futility of
lift, aerodynamic 625–626
light 29–58
Doppler shift
see Doppler shift, light
gravitational bending 581
polarization 49, 486
speed of
from Maxwell equations 515, 765–768, 1040
in substances 35
light cone 563, 566, 766
light ray 32
light-like events 563
light-year 599
line element 85–87
Cartesian 86
cylindrical 86
polar 86
spherical 87
linear hypothesis 652
linear no-threshold model 652
locality 765
loop and junction rules 848–851, 861
Lorentz group 554
Lorentz transform 517–518, 521–533, 535–555, 569, 575, 1149–1153
formula 535
inverse 536–537
linearity 1149–1153
nonrelativistic limit 533
Lorentz, H.A. 517
low-pass filter 983
LR circuit
AC 983–985
DC
current buildup 957–958
current decay 956–957
time constant 956

M

M-theory 1011, 1099
magnet 913–914
magnetic *blah blah*
see blah blah, magnetic
magnifying glass 42
malarkey 24, 197, 603
manometer 634
mass
center of
see center of mass
density, expressions for 293, 621
inertial 193
inertial vs gravitational 193, 580–581, 1039, 1041
reduced 309
mass spectrometer 917
Maxwell equations 33, 48, 515, 749–990, 1039–1041, 1078, 1079, 1085, 1087, 1092, 1094, 1095
differential form 750
integral form 750
Maxwell, James Clerk 33, 515, 666, 797, 1039
Maxwell-Boltzmann distribution

742
mechanics, definition 123
meltdown 644
meson 1070
meth, crystal
see crystal meth
method of images 812–814
metric 583–586, 1079, 1106, 1149
Robertson-Walker
see relativity, Robertson-Walker metric
Schwarzschild
see relativity, Schwarzschild metric
Michelson-Morley experiment 184, 514, 517, 1040
microscope 42–43
microwave background
see 3° background
mirror, parabolic 52–58
modern art 4
mole 677
moment of inertia 349–350, 352, 358–360, 374–380
definition 349, 350, 358, 359
table of values 374
momentum 299–302
angular
see angular momentum
conjugate 1007
conservation 300–302, 1135–1140
definition 300
operator 1052
relativistic 570–576
conservation 574–575
monopole 756, 823, 825
magnetic 756
Monty Python
see Python, Monty
motion
circular 204–206
centripetal acceleration 142, 144, 148
nonuniform 146–149
tangential acceleration 147, 148
uniform 139–145
harmonic 459–485
amplitude 459
angular frequency 459, 460
damped 477–479
definition 459
driven 479–484, 987
frequency 459
period 459
phase 460, 461
small oscillations
see small oscillations
spring oscillation 459–461
one-dimensional 129–132
acceleration, direction of 130
constant acceleration 132–135
free-fall 135
polar coordinates, in 149–153
projectile 136–139, 1113–1132
relative 132, 153
rotational 347–405
analogy with translational 347, 349, 351, 353, 354, 356
summary 403–405
three-dimensional 357–373
translational 347
two-dimensional 136
motor, electric 939
multipole 756, 825, 914, 1169
muon 1069
myopia 43

N

natural units 1030, 1078
naughty behavior 111
nearsightedness 43
neutrino 641, 1069
antielectron 639
neutron 637
Newton 3–1148
Newton (unit) 194
Newton's laws of motion 191–194
for systems 299
Noether, Emma 1140
normal force 195
potential energy of 266
work done by 258
normalization factor 668, 680
nuclear
decay 639–643
notation 638
nucleon 637
nucleus
composition 637–639
nukuler *blah blah*
see nuclear *blah blah*

nutation 402

O

octopole 825, 914, 1167–1168
Ohm (unit) 841
Ohm's bogus law 841, 842, 844–848, 868–870, 933, 938, 954, 979–982
  AC version 980
operators
  creation and annihilation 1064
  energy and momentum 1052
  ladder 1063
  raising and lowering 1063
optical instruments 42–43
orbit
  binary 432
  gravitational 209–212
  magnetic 891–892
origin, common spacetime 539
oscillations
  anharmonic 485
  harmonic
    see motion, harmonic
overtone 487

P

parabolic mirror
    see mirror, parabolic
parallel plates
  electric field of 783–784
parallel-axis theorem 379–380
paramagnetism 913
parity 359, 540, 920, 1072, 1095
particles
  real 1076
  virtual 1076, 1097
Pascal (unit) 622, 678
path integral 586, 1065–1066
pendulum 470–476
  ballistic 335
  compound 420, 474
  simple 420, 474
  small oscillations 470–474
  spherical 1027
  torsional 476
perihelia, precession 581
period 143, 459
permeability, magnetic 758
permittivity, electric 753–754, 858
perturbation theory 165, 1074
phase transition 672
photoelectric effect 1050
photon 641
physical constants, values of 19–21
pizza 253, 493
Planck's constant 21, 1050, 1051
Poincaré
  group 540, 555, 1096
  Henri 518
  invariance 1071, 1093, 1094, 1098, 1153
polar coordinates
  motion in
    see motion, polar coordinates, in
positron 1069
potential
  centrifugal 499, 501
  electrostatic 799–807
    relation to electric field 799
    relation to electrostatic potential energy 807
  function 112, 762–764
  scalar 112, 763
  vector 112, 763
potential energy
    see energy, potential
power 259–260
  rotational 354–355, 363
Poynting vector 962
precession of perihelia 581
presbyopia 44
pressure, definition 622, 678
primary (transformer) 966
principle of equivalence 581, 1041
principle of least action 1005, 1067, 1100
probability amplitude 1042, 1047, 1051, 1066, 1075
process, thermodynamic
    see thermal process
product
  cross *see* cross product
  dot *see* dot product
projectile motion
    see motion, projectile
proper time 530, 564, 570

proton 637
pulley
massive 195, 400–401
massless 195
pV plots 687, 702
Python, Monty 113, 174, 633, 1163

Q

Q (heat energy), definition 672, 685
QCD
see quantum chromodynamics
QED
see quantum electrodynamics
QF *see* quality factor
quadium 645
quadrupole 756, 799, 818, 824–825, 914, 1167
quality factor 649
quanta 1050
quantum
chromodynamics 1071
electrodynamics 1068
field theory 269, 1066–1093, 1096
mechanics 1042–1067
many-worlds interpretation 1044
tunneling 647, 1046, 1058–1059, 1104
quark 637–640, 1070–1071, 1073, 1077, 1093, 1105
quasistatic 450
process
see thermal process, quasistatic

R

rad (unit) 649
radiation, biological effects 650–652, 654, 660
radioactivity 643
radon 656–657
Rankine scale 671
Rankine, W.J.M. 671
rapidity 615–617
ray diagram 35
lens 37, 41, 42
mirror 36
RBE 649
RC circuit
AC 969, 972–974, 976–978, 980–983
DC
charging 866–868
discharging 865–866
time constant 866, 867
RC filter 983
reactor, nuclear 644–645
reduced mass
see mass, reduced
reference frame 206, 304, 511–513
center-of-mass 303–306, 308–311
inertial vs noninertial 513, 579
lab 307
reflection 32–35
law of 33
total internal 34
refraction 32–35
index of 33, 48–49
law of 33
refrigerator 694, 704, 714–717
efficiency 716–717
relativity 511–598, 749
addition of velocities 567–569
constant acceleration 588–598
electromagnetism and 768–772, 942
energy
see energy, relativistic
energy-momentum relation 574
general 517, 578–588, 1041, 1094, 1103
field equations 583–586, 1107
length contraction 558–562, 769, 770
momentum
see momentum, relativistic
pole-in-barn paradox 604
postulates 515
proper time
see proper time
rest energy 573
Robertson-Walker metric 1106
Schwarzschild metric 585, 587
simultaneity 559, 603
special 517
time dilation 555–558, 561–562, 769
gravitational 587–588
twin paradox 603

rem (unit) 649
renormalization 1075–1076
resistance 839–851
  equivalent
    parallel 846, 872
    series 845, 872
  internal 879
  power dissipation in 841–842
  resistivity, relation to 841
resistivity 840
  resistance, relation to 841
resistor 841
  impedance 978–980
resonance 482, 987
rest length 558
reversibility 717–720, 724–725, 731–732
Ricci tensor 585
Riemann curvature tensor 585
RLC circuit
  AC 985–987
road banking *see* banking
roadkill
  and differential equations 593
Robertson-Walker metric
  see relativity, Robertson-Walker metric
rockets 306–307
rolling 393–400
root-mean-square value 869–870

S

Santa Claus 722, 852
Savart, Felix 901
scalar potential
  see potential, scalar
Schrödinger equation 1042
  time-dependent 1052
  time-independent 1052
Schwarzschild metric
  see relativity, Schwarzschild metric
Schwarzschild, Karl 583
screwed
  we are all 1105, 1108
second law of thermodynamics
  see thermodynamics, second law
secondary (transformer) 966
self-inductance 953
separation of variables 214
sex 1190
Sievert (unit) 650
significant figures 16–18
silliness
  pictorial 104
  unnecessary 334
silliness of
  Aristotle 192
  chemists 670, 671
  people 17
  temperature scales 670, 671
simultaneity 559
sky, blueness of 49–51, 484
slingshot, gravitational 314
slug (unit) 195
small oscillations 464, 484–485
Snell's law
  see refraction, law of
solenoid 893, 897, 928–929
  inductance 954
  magnetic field of
    see field, magnetic, solenoidal
sound 487
  Doppler shift
    see Doppler shift, sound
  intensity 489
South Park 1036
space-like events 564
spacetime diagram 565–567, 597
special relativity
  see relativity, special
specific heat 672–673
spectrum, electromagnetic 30–32
speed 126
spontaneous symmetry breaking 1071–1072, 1103, 1105
spring
  constant 268
  force law 268–269, 460
  harmonic 268
  vertical 468–470
stake, burning at the 674, 1039
Standard Model 1067–1077, 1095–1099
state
  equation of 677, 702
  nonequilibrium 686, 703
  variable 677–678, 686, 696, 702
static equilibrium
  see equilibrium, static
statics
  see equilibrium, static
statistical fluctuations, thermal

665–666, 668, 683
statistical mechanics 666
Stokes's theorem 100–104, 893
string theory 7, 555, 1011, 1097–1103
strong interaction 638, 1070, 1073, 1075, 1077, 1093, 1095, 1103
summation convention 583, 1079, 1136, 1149
sunrise and sunset, redness of 51–52
superconductor 812, 840
supergravity 1096–1097
superposition
  center of mass 296–297
  principle of 761–762, 783–784
supersymmetry 1096–1097
surface element 90–92
  Cartesian 90
  cylindrical 91
  polar 91
  spherical 91
symmetry 191, 253, 538–549, 586, 773, 777, 787–790, 792, 1011, 1041, 1067–1072, 1078–1103, 1140
  cylindrical 778
  gauge 764–765, 1069–1071, 1078–1093
  global 1084
  local 1084
  planar 780
  spherical 774
system 193
  momentum of 301
system, thermally isolated
  see thermal isolation
syzygy 178, 406

T

tau 1069
tautochrone 1002–1004
telescope 42–43
temperature 670–671
  conversions 671
  measurement of 714
  versus energy 670
tension 195
terminal velocity
  see velocity, terminal
Tesla (unit) 759
thermal
  cycle 687
  equilibrium 674, 703
  isolation 674, 676
  process 686, 702
    adiabatic 697–703
    isobaric 691–692, 703
    isochoric 689, 703
    isothermal 689, 702
    quasistatic 703
thermodynamics 666
  first law 685–686, 702
  second law 720–722, 724–725, 731
    statistical nature 728–729
thought experiment 558
timbre 487
time dilation
  see relativity, time dilation
time reversal 1072
time, proper
  see proper time
time-like events 563
torque 350–353, 360–362
  arm 352
  definition 350, 351, 361
  gravitational 392–393
tortoise, the 605
transformer 966
transient 246, 483, 978
transients 977–978
triangle trick 792
tritium 645
Tryon, Edward P. 1035
tunneling, quantum
  see quantum tunneling
turbulence 627–628
twin paradox 557

U

u (unit) 608, 638
uncertainty principles 1047, 1061
Understanding, The Great Onion of 177
unification 191, 1067, 1068, 1071–1072, 1077, 1094–1102
units
  conversions 19
  electromagnetic 751

universal gravitational constant 208, 1107
uranium, separation of isotopes 645

V

vacuum bubble 1076, 1102, 1103
vacuum, false
see false vacuum
Van de Graf generator 817
variations, calculus of
see calculus of variations
vector 65–78
addition 65
analytic 68–69
axial 74
components 67
scalar 70
vector 70
inverse 67
position 85, 123
subtraction 67
unit 69–70
vector potential
see potential, vector
velocity 123, 125
angular
see angular velocity
average 125
escape 281
generalized 1005
instantaneous 125
relativistic addition of
see relativity, addition of velocities
terminal 246
Venturi tube 634
virtue, disregard for 3
voltage
see potential, electrostatic
volume element 92
Cartesian 92
cylindrical 92
spherical 92
Voynich manuscipt 750

W

$W^{\pm}$ 1069
W^{-} 639
Watt (unit) 259
wave
amplitude 29, 768
equation 767
frequency 30, 35, 578, 768, 1030, 1048, 1050
function 1042
interference 45–48
light 29–33, 45–49, 486, 515, 1040, 1050
longitudinal 487
number 768, 1030, 1048
period 30, 768, 1048
plane 29, 1030, 1051
propagation 486
sound *see* sound
speed of propagation 30, 35, 768, 1049
standing 487
superposition 45–48
transverse 29, 486
traveling 29, 486–487, 767, 1047
wavelength 29, 31, 35, 490–491, 578, 768, 1048
weak interaction 1069, 1071–1073, 1075, 1077, 1093, 1095, 1103, 1105
weight 195
perceived 212–213
vs mass 195
weirdness and perversion 364–366
Wheatstone bridge 880
Wheeler, John 583
will to live, loss of 463
work 253–271
bogosity of 253, 254
definition 255, 257
done by a constant force 257–258
done by magnetic force 891, 951
done by normal force 258
higher-dimensional 256–257
one-dimensional 254–256
rotational 354, 363
work-energy relation 256

Y

yada, yada, yada 153, 211, 702

Z

Z 1069
Zen 7–11, 851